INTEGRATING STATISTICAL AND MECHANISTIC MODELING TO ANALYZE DISEASE OMIC DATA

BY

YULIANG WANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Chemical Engineering
in the Graduate College of the
University of Illinois at Urbana Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Professor Nathan Price, Chair
Professor Huimin Zhao
Associate Professor Christopher Rao
Assistant Professor Jian Ma

Abstract

The advent of high throughput technologies has enabled large-scale measurements of the genome, transcriptome, proteome and metabolome of tissues samples, serum and even single cells. Additionally, prior biological knowledge is increasingly curated into accessible databases and reconstructed into computable models. My research aims to integrate high throughput data and prior knowledge to improve disease diagnosis and our understanding of biological systems, by leveraging the power of both statistical learning and mechanistic modeling approaches.

The first part of my Ph.D. work is to apply increasingly mechanistic biological constraints in in the analysis of high throughput gene expression data to identify molecular signatures of disease phenotypes. Chapter 2 discusses the statistical issues and recommended steps to generate accurate and reproducible molecular signatures. Chapter 3 presents a new computational method that uses the relative expression level of interacting gene pairs as accurate molecular signatures. By incorporating prior knowledge about the relations between genes, this method increases molecular signature reproducibility compared with previous methods.

Metabolic networks reconstructed from known reaction stoichiometry and gene-protein-reaction associations provide a mechanistic context to analyze gene expression data. In Chapter 4, I developed a new analysis pipeline that identified perturbations at metabolic branch points (i.e., structures where two reactions consume the same metabolite). Different phenotypes (e.g., cancer v.s. normal) can be accurately distinguished by transcriptional changes at metabolic branch points. Combining reaction expression state (high/low), mass conservation and thermodynamic constraints, I identified additional perturbed branch point reaction pairs that are not apparent from expression data alone.

The second part of my PhD work is to contextualize and refine prior knowledge by integration with context-specific high throughput data. In Chapter 5, I developed a novel computational method mCADRE to reconstruct tissue-specific metabolic models. This method can use transcriptomic,

proteomic and metabolomics data to infer the metabolic network of a given tissue or cell type. This method can be viewed as using tissue-specific omic data to refine and contextualize prior knowledge of metabolism. Using this new method, I reconstructed genome-scale metabolic models for 126 human tissues, providing a tissue-specific encyclopedia of metabolism. In Chapter 6, I applied mCADRE to reconstruct metabolic networks of commonly used breast cancer cell lines. Systematic comparison of model prediction and experimental results revealed different types of inconsistencies that call for further model curation and the development of new modeling approaches.

Acknowledgements

In an era where "Big Data" in biomedical sciences holds great promise to transform our understanding of complex biological systems and our approach to healthcare, I feel very fortunate to have chosen computational and systems biology as my doctoral field of study. In the past four years, I have grown immensely as a researcher, and I want to acknowledge many great individuals that have been integral to my success.

First and foremost, I will always be grateful to my advisor, Nathan Price. When choosing research direction at the beginning of graduate school, Nathan introduced me to computational biology, where I can combine my deep interest in biology and my strong quantitative skills as an engineer. Since I joined the lab, Nathan has encouraged me to develop and pursue research ideas independently, and engaged me in thoughtful discussions at critical moments. I also benefitted a lot from Nathan's focus on the big picture and novel, ambitious questions. Nathan is also instrumental in creating a lab environment where we are always happy to help and learn from each other, and I owe a significant part of my growth to this constructive environment.

Besides Nathan, I am fortunate to work with Prof.Don Geman from The Johns Hopkins University. I enjoyed his insightful questions and strong statistical perspectives. I would also like to thank the remaining members of my committee, Huimin Zhao, Christopher Rao, and Jian Ma. Their insightful questions and feedback during my preliminary and final exams provided valuable guidance.

I owe my four happy years at the Price lab to many good friends and lab mates, especially Matthew Benedict, Shuyi Ma, Jaeyun Sung, Andrew Magis, Sriram Chandrasekaran, James Eddy and Chunjing Wang. I thoroughly enjoyed my time with them in both Champaign-Urbana and Seattle. Over the years, I got lots of invaluable help, encouragement and advice from them.

Last but not least, I want to thank my family: my parents, Qingxiang Wang and Ruzhen Shen, my sister Meiying Wang and my girlfriend Peiyun Zhou. Peiyun always believes in me, full of confident in my potential and encourages me at moments of doubt. My parents never went to college and never travel far from the village they were born into. Yet they have the vision that education is the key to a better life and worked extremely hard to make sure that their son went to one of the best universities in China. They were also very supportive when I decided to pursue graduate study in the United States, even though all they knew was that U.S. is far way country excellent in science and technology.

Table of Contents

Chapter 1. Introduction and Overview

Decades of molecular and cellular biology research have generated enormous knowledge about how biological systems work from molecular to physiological levels. Such prior knowledge is increasingly curated in easily accessible databases [1, 2]. On the other hand, the advent of high throughput molecular profiling technologies, most notably gene expression microarrays [3], enable the comprehensive measurement of a large number of molecular entities simultaneously. Capitalizing on advances in various "omics" technologies, large community projects such as The Cancer Genome Atlas (TCGA) have generated enormous amounts of data that characterize hundreds of diverse tumor samples with the genome sequence, epigenetic changes (methylation), gene expression and DNA copy number [4-9]. Similar efforts have been devoted to commonly used cancer cell lines [10-12]. The key challenge is to gain novel information from both prior knowledge and a large amount of omic data that can aid disease diagnosis and improve our understanding of biological systems.

Herein, I present new computational tools and methodologies I developed that aim to combine prior knowledge with high throughput data to guide disease diagnosis and generate new hypotheses. The development of these tools uses both statistical learning approaches and mechanistic modeling. The goal of this chapter is to provide a high-level introduction to both approaches and an overview of chapters in the thesis.

1.1 Learning Disease Molecular Signatures from High Throughput Data

A molecular signature is defined as "a set of biomolecular features (e.g. DNA sequence, DNA copy number, RNA, protein, and metabolite expression) together with a predefined computational procedure that applies those features to predict a phenotype of clinical interest on a previously unseen patient sample" [13]. In the most typically case of transcriptomics, the mRNA level of p genes are measured in n labeled samples (binary: disease/ normal or continuous: survival time). The aim is to use various statistical approaches to construct a computational model consisting of a subset of

informative genes that accurately predict the defined endpoint. Molecular signatures have been developed for disease diagnosis (e.g., disease v.s. normal), prognosis (e.g., long v.s. short survival), and treatment selection (e.g., responsive v.s. resistant). Molecular signatures derived from gene expression profiling such on MammaPrint [14] and Oncotype DX [15] have been used in clinic to assess risk of breast cancer metastasis and recurrence.

During the past decade, numerous tools have been developed to derive molecular signatures from (mainly) gene expression data. Initially, standard statistical and machine learning approaches were applied to gene expression data. This includes using statistical tests (t-test or Wilcoxon test) to identify differentially expressed genes (DEGs) between two conditions, or use general-purpose classification methods such as Support Vector Machines (SVM) and k-nearest neighbors (kNN) to classify samples. Tools specifically tuned for the "large *p*, small *n* situation" in gene expression data analysis, such as Significant Analysis of Microarrays (SAM)[16] and Prediction Analysis of Microarrays (PAM)[17] were developed. These methods often ignore the dependence structure between genes.

Newer methods increasingly aim to put more biological constraints on gene expression data analysis, beginning with methods that aim to identify biologically meaningful gene sets that show coordinated expression changes between conditions. Gene Set Enrichment Analysis (GSEA) [18] is the most popular approach among them. The aim of GSEA is to determine whether members of a pre-defined gene set S (biological pathways) tend to occur toward the top (or bottom) of the list L of differentially expressed genes. GSEA ignores the connections between genes in biological pathways. Later on, methods that account for biological pathway structure, such as impact factor (IF) analysis [19] were developed. To extend analysis beyond existing pathways, which only cover a small fraction of well-studied human genes, Chuang *et al* [20] developed a method that identify modules enriched with differentially expressed genes in a protein-protein interaction network to classify breast cancer patients. Therefore, the field of molecular signature discovery has been evolving from single genes, to a priori defined gene sets, to network-based diagnostic signatures.

1.2 Genome-Scale Models of Metabolism

Metabolism is fundamental to all cellular processes. Intermediates in metabolism are intimately linked to cellular signaling: ATP is the substrate for phosphorylation in kinase cascades; acetyl-CoA is the substrate of acetylation of histones that alter chromatin dynamics; S-adenosyl methionine provides substrate for DNA methylation. On the other hand, metabolism is under extensive control of cellular signaling, with PI3K/Akt/mTOR pathway being the most prominent regulator. Perturbations in metabolism are found in most human disease such as inborn errors of metabolism, cancer, diabetes and obesity, and neurodegenerative diseases.

Metabolic networks are among the best studied biological networks, thanks to decades of detailed biochemical experiments. Such prior knowledge is systematically curated into *computable* genome-scale metabolic models (e.g., Recon 1 [21] and Recon2 [22]). The latest human metabolic network, Recon 2, includes over 7000 metabolic reactions, 2600 metabolites, and 1700 metabolic genes[22]. The content of a metabolic network can be divided into two parts. The most basic part is the stoichiometric matrix, where rows represent metabolites, columns represent reactions, and numeric entries (i, j) represent the stoichiometric coefficient of metabolite i in reaction j. The other part is the gene-reaction rules, which describe what genes encode what proteins and how these proteins are organized to catalyze reactions (isozymes or enzyme complexes). Various constraints can be applied to the metabolic network to reduce the number of possible metabolic states under a particular condition. Typical constraints include mass balance, thermodynamic constraints, nutrient uptake rates, expression of metabolic genes (e.g., penalize flux through reactions catalyzed by lowly-expressed enzymes). The reconstruction and simulation of genome-scale metabolic networks are well-established processes and there are many computational tools to explore the capabilities of metabolic models [23].

There are numerous different types of human tissues (e.g., liver, lung, breast) and cell types (e.g., Myocytes, adipocytes, hepatocytes). Different tissues express different sets of metabolic genes and carry out a different subset of all metabolic capabilities encoded in the genome: gluconeogenesis and urea cycle occurs exclusively in the liver, while lipid synthesis and transport genes are highly

enriched in adipose tissues. A few computational methods have been developed to account for such tissue-specificity in metabolism [24-26].

One important application of tissue-specific metabolic models is to study metabolic aberrations in cancer and identify selective metabolic targets. Metabolic reprogramming to fuel proliferation is a hallmark of cancer. Well-known oncogenes and tumor suppressor genes regulate different aspects of metabolism[27]. Mutations in metabolic genes (e.g., SDH, FH, IDH) are also causally involved in tumorigenesis [28]. Common metabolic aberrations shared by most types of cancer are increased glucose uptake, increased lactate secretion and greater tendency to use glutamine. However, depending on tissue context and underlying genetic lesion, cancer tissues can also have prominent metabolic differences [29]. This heterogeneity is also reflected at the transcriptomic level: while upregulation of nucleotide biosynthesis and glycolysis are frequently observed across tumors, expression changes of other pathways (e.g., oxidative phosphorylation) are very heterogeneous [30]. Therefore, it is not only important to develop metabolic models that represent the common features of all cancer types, but also cancer type specific metabolic models that represent the unique metabolic capabilities of tumors with different tissues types and genetic lesions. Such tissue-specific metabolic models have been used to identify metabolic targets that selectively affect cancer proliferation. In particular, using metabolic modeling, it was identified that haeme oxygenase is synthetically lethal with FH, which is frequently mutated in renal carcinoma. Therefore, targeting FH would only affect RCC with mutant FH but spare normal cell with wild type FH [31]. As genome sequencing continues to identify both copy number loss and putative damaging mutations in metabolic genes, using genome-scale metabolic models to systematically identify synthetic lethal partners of cancer-specific metabolic mutations might be a promising path to selective drug targets.

1.3 Dissertation Organization

Chapter 1 introduces the central themes and topics for the work described in subsequent chapters.

Chapter 2 systematically reviews challenges and consensus in generating molecular signatures from omic data.

Chapter 3 describes a new computational method that combines prior information on protein-protein interaction with gene expression data to identify biologically interpretable interactions that accurately classify clinically relevant disease phenotypes. This is a statistical modeling approach built upon an existing family of methods called Relative Expression Analysis.

Chapter 4 describes a new computational tool (metabolic Context Assessed by Deterministic Reaction Evaluation, mCADRE) to reconstruct tissue-specific genome-scale metabolic networks based on prior biochemical knowledge and tissue-specific transcriptomic, proteomic and metabolomic data.

Chapter 5 describes the application of mCADRE to build genome-scale metabolic models of commonly used breast cancer cell lines and the analysis of various functional genomic data using the cell line specific metabolic models.

Chapter 6 describes the analysis of transcriptomic data in a metabolic network context, which considers both network topology and function.

Chapter 7 summarizes the dissertation and provides perspectives on future developments.

Chapter 2. Molecular signatures from omics data: from chaos to consensus

2.1 Introduction

In recent years, new high-throughput measurement technologies for biomolecules such as DNA, RNA, and proteins have enabled unprecedented views of biological systems at the molecular level. The fields of research associated with obtaining and understanding such measurements – for instance, genomics, transcriptomics, and proteomics – are sometimes referred to in aggregate as *omics*. Given molecular measurements taken from a biological system, a natural goal is to develop a statistical model that uses these measurements to predict a clinical outcome of interest, such as disease status, survival time, or response to therapy. In this article, we will discuss the process of using omics data to discover a *molecular signature*. Here we define a molecular signature as *a set of biomolecular features (e.g. DNA sequence, DNA copy number, RNA, protein, and metabolite expression) together with a predefined computational procedure that applies those features to predict a phenotype of clinical interest on a previously unseen patient sample.* A signature can be based on a single data type [3, 32-34] or on multiple data types [35-38]. The overall process of identifying molecular signatures from various omics data types for a number of clinical applications is summarized in Figure 2.1.

Many possible clinical phenotypes might be predicted by a molecular signature; a few examples include prediction of disease risk and progression [39-41], response to therapeutic drugs [42-44] and their physiological toxicity [45, 46], and time to disease recurrence or death [47, 48]. (Note that in this review, the molecular signatures that we consider may be effect modifiers or may only be of prognostic value; in either case, we refer to the molecular signature as "predicting" a clinical phenotype of interest.) A successful case of the clinical utility of omics-derived molecular signatures is MammaPrint [14], a diagnostic test approved by the Food and Drug Administration for clinical use. MammaPrint is a 70-gene expression signature used to predict breast cancer prognosis and to determine the appropriate therapeutic regimen for lymph node negative breast cancer patients with either ER positive or negative. The list of 70 genes was selected based on correlation with clinical

outcome (distant metastasis vs. no metastasis), and underwent successful validations on independent patient cohorts [49, 50].

Despite a few notable exceptions such as MammaPrint, the successful discovery of molecular signatures has largely been hampered by limited reproducibility and variable performance on independent test sets [51-57], as well as difficulty in identifying signatures that outperform standard clinical measurements like the cardiovascular disease risk C-reactive protein (CRP) [58]. These difficulties can be attributed in large part to the low signal-to-noise ratio inherent to omics datasets, the prevalence of batch effects in omics data, and molecular heterogeneity between samples and within populations [59]. These issues are exacerbated by the fact that the datasets used to develop molecular signatures tend to have small sample sizes relative to the number of molecular measurements [60]. Moreover, improper study design, inconsistent experimental techniques, and flawed data analysis can lead to further challenges in the process of molecular signature discovery. Though there has been marked progress in the field of molecular signature discovery in recent years, there remains a clear need for further improvements in the discovery process in order for omics-based technologies to begin to achieve their full clinical potential.

2.2 The four stages of molecular signature discovery

Roughly speaking, the process of molecular signature discovery on the basis of omics data consists of four major stages:

1. Defining the scientific and clinical context for the molecular signature.
2. Procuring the data.
3. Performing feature selection and model building.
4. Evaluating the molecular signature on independent data sets.

In the sections that follow, we will discuss each of these stages in turn.

Stage 1: Defining the scientific and clinical context

Before embarking on the process of molecular signature discovery, one must first identify a specific scientific and clinical context for the molecular signature. A molecular signature uses omics measurements to predict a clinical phenotype of interest; therefore, it is natural that before constructing such a signature, one must first determine what type of omics measurements will be used, and what clinical phenotype will be predicted.

We first consider the problem of selecting a suitable omics data type for a molecular signature. A signature intended to distinguish between cancer and normal tissue could be based upon a number of omics data types; for instance, one might base the signature upon gene expression measurements, if it is believed that this type of cancer shows altered expression of some genes relative to normal tissue, or upon DNA sequence data, if samples from this cancer are characterized by particular mutations or copy number changes. However, given a clinical phenotype of interest, certain types of omics data might not form the basis for a sensible molecular signature. For instance, it would not be reasonable to attempt to create a molecular signature to screen for adult onset (type II) diabetes on the basis of DNA sequence data alone because an individual's DNA sequence remains essentially static throughout his or her lifetime, but risk of developing the disease may change.

We now consider the clinical context of the molecular signature. A gene expression-based signature that can distinguish between cancer and normal tissues would be of little practical use if a physician can easily make the same distinction using standard (and less expensive) clinical approaches. Similarly, a signature that can distinguish between two subtypes of cancer is useful only if those two subtypes differ in some clinically relevant way, such as in survival time or response to therapy, since otherwise the information about cancer subtype provided by the molecular signature may not serve a practical purpose. As an example, gastrointestinal stromal tumors (GISTs) and leiomyosarcomas (LMSs) are remarkably similar morphologically and were originally classified as being the same cancer. However, it was found that they respond very differently to distinct therapies, and thus a signature that can distinguish between these two diseases based on gene expression in tissue samples

can be useful [34]. An example outside of cancer involves the use of metabolomic information from human serum to noninvasively diagnose and monitor Alzheimer's Disease (AD) progression [61-63].

Stage 2: Data procurement

The development of a molecular signature requires the availability of adequate omics data for which the clinical phenotype of interest is available. In general, there are two ways in which such data can be procured: new data can be collected experimentally for the specific purpose of molecular signature discovery, or else existing data (collected previously for other purposes, and generally publicly available) can be used. There are pros and cons of either approach. Collecting new data has a major advantage, in that all aspects of the experiment can be carefully controlled. On the other hand, data collection is expensive, and given the large sample sizes necessary for successful molecular signature discovery, using existing data sets may be a more feasible approach. There are a number of public data repositories from which omics data and associated clinical phenotypes can be obtained. For instance, a useful source of gene expression data is NCBI Gene Expression Omnibus (GEO), a repository of over twenty six thousand studies that continues to grow at a rapid pace. Other public data repositories include ArrayExpress [64] and Sequence Read Archive [65]. Regardless of how the data are procured, it is crucial that the samples correspond to the scientific and clinical context of interest, as described in the previous section.

In order for a data set to be suitable for molecular signature discovery, the samples must be collected under appropriate experimental and analytical conditions. As an example, any biological factors (such as gender, age, or ethnicity) that may be associated with the clinical phenotype of interest or with the omics measurements should be taken into consideration in the process of data procurement. In addition, to reduce the prevalence of *batch effects*, factors such as sample collection and processing procedures, laboratory personnel, study run-dates, reagent sources, measurement instruments, and data processing methods should be carefully controlled [66-68]. Deviations in these protocols can have a surprisingly large effect on the omics measurements obtained, often larger than the effect of the clinical phenotype of interest [69]. Ideally, there should be no association between the clinical phenotype of interest and these factors. For instance, in the case of a molecular signature that

classifies tissue samples into tumor versus normal, there should be no difference between the tumor and normal samples in terms of the laboratory personnel who performed the sample preparation, or the sample run-dates. If experimental and analytical procedures are not carefully controlled, they can result in confounding with the clinical phenotype of interest, leading to the development of a classifier that performs very well on the data used in its development, but that will perform poorly on independent test samples.

To the extent that analytical and experimental factors do vary among the samples, these factors should be explicitly included in the model used to develop the classifier. Normalization procedures have been proposed that are intended to reduce the effect of measured and unmeasured external factors on omics data [70]; however, good experimental design remains the best strategy [71]. Exploratory data analysis techniques, such as hierarchical clustering (Fig. 2.2A) and principal components analysis (Fig. 2.2B) can be useful tools to assess the extent to which covariates that are not of primary interest may have affected the data.

When existing data is used for omics-based molecular signature discovery, it is particularly important that sufficient information about the experiment is available to ensure that good experimental design was followed (this will be discussed further in Section 4). For instance, if the run date for each sample is not given, then one cannot be certain that the clinical phenotype of interest is not highly confounded with run date.

Unfortunately, many omics studies have sample sizes substantially smaller than would be required for the successful identification of molecular signatures. A molecular signature that is developed on the basis of a small number of samples is more likely to be sensitive to technical and biological sources of noise and variation, and less likely to capture the aspects of the data that are truly associated with the phenotype of interest. This exacerbates the risk of over-fitting, wherein the signature performs well on the samples used for signature development but fails to correctly predict the clinical phenotype of interest in previously unseen samples. In contrast, global molecular characteristics of a particular phenotype may become more apparent as sample size increases. Therefore, having a large sample size, while by no means a cure-all, will greatly improve the odds that a given attempt at molecular

signature discovery will prove fruitful. Integrating across multiple datasets of the same phenotypes from different labs can also help to amplify the primary biological signal of interest relative to noise. Of course, whether a given sample size is "large" or "small" depends the type of omics data being used for signature discovery, the clinical phenotype of interest, and many other factors.

Stage 3: Feature selection and model building

Once a scientific and clinical context has been established and one or more data sets have been identified, we can develop a molecular signature through (1) feature selection; and (2) model building. These two tasks can be performed together or separately.

We first consider the task of feature selection. A typical omics experiment simultaneously measures thousands or even millions of biological features (e.g. single nucleotide polymorphisms, RNA transcripts, protein levels) on each patient sample. However, just because thousands of molecular measurements are obtained does not mean that thousands of molecular measurements should be used in the molecular signature. Since financial cost, technical practicality, and measurement robustness are important criteria to select signatures, then if all else is equal, a signature that could be ultimately measured via PCR or Western blot is favored over a signature that requires a technique involving many more protocol steps, such as in omics measurements. In order to reduce the number of features used in molecular signature development, *feature selection* is performed. Feature selection can be performed in a *supervised* manner (e.g. the 20% of features that are most associated with the clinical phenotype of interest are selected), or in an *unsupervised* manner (e.g. the 20% of features with the highest variance are selected). Once a set of features has been selected, only those features are used in the model building process, which is described next.

We now consider the task of *model building* – that is, the process of developing a specific computational procedure that can be applied to the omics measurements from a future patient sample in order to predict the unknown clinical phenotype of interest for that sample. There are many possible approaches to building such a model, and in particular, the type of model used will depend on the clinical phenotype of interest. For instance, if we wish to develop a molecular signature to predict

time to cancer recurrence, then a Cox proportional hazards model might be appropriate. On the other hand, to develop a molecular signature that can distinguish between cancer and normal tissue, one could use a classification approach, such as logistic regression, support vector machines, neural networks, or linear discriminant analysis. Some approaches for model-building involve first performing an unsupervised technique, such as clustering or principal components analysis, followed by a supervised procedure, such as logistic regression.

It is worth noting that it is not always obvious what type of model should be used in a given setting. For instance, suppose that we wish to develop an expression-based signature in order to distinguish between tumor and normal samples. It sounds easy enough. An obvious approach is to develop a binary classifier, using e.g. logistic regression, which assumes that there is a linear boundary separating the two classes (tumor and normal). However, in some settings, this assumption might not be appropriate. For instance, maybe the normal samples do not belong to a single homogeneous group: there may be differences among the normal patients that are at least as great as the differences between tumor and normal patients. Alternatively, perhaps the tumor samples are heterogeneous because there are in fact several distinct subtypes of the tumor. In such a setting, a binary treatment of the problem that assumes a linear decision boundary may be inappropriate. Therefore, it is important to choose a model that is well-suited for the scientific and clinical contexts.

Once we have developed a model, how can we determine whether it is any good? Despite certain drawbacks [72, 73], the most popular approach for evaluating model performance in this context is *cross-validation.* (Cross-validation is also often used for tuning parameter selection, though that application is outside of the scope of this article.) Cross-validation involves repeatedly splitting the samples in the data set into training and test sets, performing all aspects of feature selection and model building on the training set, and evaluating the model's performance on the test set. Cross-validation can also be used to select from among a small number of possible models: the model with the smallest cross-validation error rate should be chosen.

Cross-validation is a simple and intuitive approach to estimating the error rate associated with a model, but it must be performed with care. Most importantly, within each cross-validation fold, no

information about the test set can be used in building the model on the training set. For instance, suppose that one performs feature selection by selecting the 10% of features whose t-statistics between cases and controls are largest. One then performs logistic regression, using only these features, to develop a classifier to distinguish between cases and controls. How should the cross-validation error rate be calculated? Consider the following two approaches:

Approach 1 (incorrect)

> Identify the 10% of features that differ most between cases and controls, and use only those features henceforth. Perform cross-validation by repeatedly splitting the samples into training and test sets, fitting a logistic regression model on the training set (using just the 10% of features previously identified), and then evaluating the model's performance on the test set.

Approach 2 (correct)

> Perform cross-validation by repeatedly splitting the samples into a training set and a test set. Within each training set, identify the 10% of features that differ most between cases and controls, and use those features to fit a logistic regression model. Then evaluate the performance of this model on the test set.

The difference may seem subtle, but it is in fact crucial. Approach 1 will yield a woeful underestimate of the true error rate, because the 10% of features that differ most between cases and controls were identified using all of the samples, including those in the test set, rather than simply the training samples. In effect, if Approach 1 for cross-validation is taken, then perfect error rates can potentially be obtained even on data sets in which the "case" and "control" labels were assigned randomly! On the other hand, in Approach 2, feature selection is performed using the training set within each cross validation fold, and so the resulting cross-validation error rate is valid. Unfortunately, the difference between Approaches 1 and 2 is often overlooked, and the literature is rife with papers in which extraordinarily low, but grossly inaccurate, cross-validation error rates are reported because some variant of Approach 1 has been performed. The key principle is that in computing cross-validation error rates, within each cross-validation fold only training observations can be used in any aspect of

feature selection or model development. Deviations from this principle, even if seemingly innocuous, may result in dramatic underestimates of error.

At the end of the feature selection and model building process, the molecular signature must be *locked down* – that is, the precise computational procedure used to convert a new omics sample into a prediction of the clinical phenotype must be completely specified. Only then can the molecular signature be fairly evaluated on independent data sets, as described next.

Stage 4: Evaluation on independent data sets

Once a promising molecular signature has been identified, its performance needs to be evaluated on completely independent patient samples. Unlike cross-validation, wherein the test set is drawn from the same population as that of the training set, an *independent* sample is one that is completely separate from the set of samples used for feature selection and model building. In particular, this means that the test set is *not* simply a random split from a large dataset (even if sequestered and not used in any training sets). If a molecular signature performs well on a truly independent set of samples, then this provides evidence that it will likely generalize to future patient samples. However, the amount of evidence for a molecular signature's performance based on independent data depends critically upon specific characteristics of the independent data set.

Lower level of evidence. Good performance on an independent data set collected at the same institution using carefully controlled protocols. This provides evidence that the molecular signature works well in this particular setting, with these protocols, with the patient profile at this institution, etc. However, it may not hold up elsewhere. At the very least, its ability to work in other settings has not been demonstrated.

Higher level of evidence. Good performance on multiple independent data sets collected at multiple institutions. Success in this setting is the best evidence that a molecular signature will perform well on future patient samples. This indicates that the signature is robust to the kinds of things that might

change between locations: namely, aspects of the biology of the populations that tend to go to particular hospital, sample preparation and measurement techniques used, and so forth.

Evaluation of a molecular signature on fully independent patient samples is the gold standard for assessing its performance. Unfortunately, it often is the case that molecular signatures that seem promising in the feature selection and model building stage (i.e. that have very low cross-validation error rates) exhibit poor performance on independent data.

2.3 Disclosing all experimental protocols, data sets, and source code

A key principle of science is that other researchers must be able to reproduce the results. In order for a molecular signature to be reproduced, three essential pieces of information are required: 1. The experimental and analytical protocols; 2. The raw data; and 3. The source code used to develop the signature. We discuss each of these points in turn.

In order for a molecular signature to be fully understood by other researchers, detailed information on the experimental protocol, including the patient selection criteria and experimental and analytic procedures, must be made available. Without this information, one cannot determine the scientific or clinical contexts in which the molecular signature is intended, appropriate, or useful.

Second, in order for a molecular signature to be reproduced, the omics data used in its development, as well as the associated metadata and clinical data, must be made available. If the data are not released, then it simply is not possible for other research groups to determine whether the molecular signature is valid. Since large sample sizes are generally required in order to develop satisfactory molecular signatures, it is infeasible due to both time and cost constraints for another group to collect their own data set in order to validate the molecular signature. In addition to allowing for independent confirmation of the molecular signature (and thereby increasing confidence in its scientific merit), releasing data also serves to further science, since then other investigators can use the data for their own molecular signature development. This is particularly important because in many applications, no

single research group will be able to collect a sufficiently large data set, making meta-analyses of large numbers of published datasets highly valuable as an alterative approach. Given the large public investment in biomedical science, there is a strong argument for omics data to be made publicly available whenever possible, so that it can be leveraged maximally for the public good [74, 75].

Finally, even if the data are made available, other research groups will not be able re-derive the molecular signature based on the same data used for its discovery, and confirm that the signature does truly work well on independent data, unless all data processing techniques and all analytical and computational methods are made available. Unfortunately, in practice this information often is not provided in sufficient detail. For instance, there is a tendency for authors to publish a list of the features (e.g. genes) involved in the signature, without the detailed mathematical formulas required to understand precisely how the omics measurements are used in order to predict the clinical phenotype of interest. This is a major obstacle to progress in the field, as other research groups cannot reproduce or validate – much less build upon – research that is not sufficiently reported. Unfortunately, due to the complexity of omics data sets and the analyses required to develop molecular signatures, it is almost impossible to describe an analysis in sufficient detail that another researcher could exactly reproduce those steps. In order to address this problem, the source code used to develop the molecular signature should be released. Ideally, this code should encompass all aspects of signature development, from processing and normalization of the raw omics data, to feature selection to model building to evaluation on an independent data set.

2.4 Using multiple data sets for molecular signature discovery

Thus far, we have described the development of a molecular signature on the basis of a single data set, followed by evaluation of the signature on one or more independent data sets. However, in principle, multiple data sets can be used for molecular signature discovery. In fact, this can often lead to more accurate and more broadly applicable molecular signatures.

When a molecular signature is developed on the basis of a single data set and then tested on an independent data set, its performance tends to degrade severely in the independent data set relative to

its cross-validation error rate in the data set used for development. This drop in performance can stem from heterogeneity between studies due to underlying variance in the biology of the patients studied, as well as from technical variations in measurement, normalization, and analysis. That is, a signature developed using a single data set may overfit certain aspects of the data set that are not of primary scientific interest, leading to poor performance on independent data. This problem can be partially overcome by developing the signature on the basis of multiple data sets, collected at different institutions and at different time points [76-78]. (However, the primary clinical phenotype of interest, such as tumor versus normal, must be balanced between the data sets in order to avoid confounding between the data sets and the clinical phenotype.)

2.5 Using multiple data types for molecular signature discovery

Given the complexity of biological systems in general and pathological processes in particular, there is an upper limit to how well a molecular signature developed on the basis of a single data type (e.g. genome-wide expression on DNA microarrays) can predict disease phenotypes and clinical outcomes. Integrating multiple types of omics data may allow for the development of increasingly accurate and robust molecular signatures. For example, gene expression data can be combined with copy number variation data or DNA sequence data. Successful multi-scale integration of different types of biological information is one of the current challenges in systems biology [79, 80]. In **Fig. 2.3**, we provide brief summaries of a few recently published studies [79-86] in which multiple data types were used for molecular signature discovery.

A number of methods to combine diverse types of omics data across different measurement platforms and laboratories have been proposed [79, 80, 87], in order to more accurately select clinically relevant features or to develop better molecular signatures. For example, English and Butte evaluated data from 49 obesity-related studies that used different experiment types, including DNA microarrays, genome-wide association, proteomics, and RNAi knockdowns [82]. The investigators found that the biomolecules reported to be associated with obesity in individual studies had little overlap with previously known obesity-related genes. The investigators then determined a gene to be obesity-related if 5 or more studies reported the gene to be obesity-related. Using this approach of

feature selection, they were able to identify a higher proportion of known obesity related genes than from any of the 49 individual studies, and also discovered new genes for which there was compelling support of association with obesity [82]. This demonstrated that even straightforward integration of multiple omics data types can substantially improve the feature selection process. In a study by Lu *et al.,* the investigators integrated data types in order to perform more effective feature selection: they identified 475 genes that were differentially expressed between lung adenocarcinoma and normal tissue, and that were also located in copy number varying regions [83]. This gene set was used to create a predictive model for patient survival, which was then shown to be accurate on three independent patient cohorts. Advances in integrating diverse omics data types may lead to a reduction in spurious signal caused by technical limitations of individual platforms, and an increased ability to identify molecular signatures associated with the underlying mechanistic roles in disease pathogenesis.

2.6 A network-based approach to molecular signature discovery

The use of network-based approaches is a promising avenue for molecular signature discovery. These networks represent a complex web of interactions among diverse components in a cell, and can be used to develop more reproducible and accurate molecular signatures by exploiting the underlying biology of the system. Network-based approaches extend beyond simple integration of different omics data types, and can involve evaluating complex interactions that can vary due to disease or other perturbations.

Most statistical methods for feature selection and model building do not take a network-based approach: they implicitly assume that the features are independent, or that they are only weakly dependent, though this has begun to change in recent years [88-90]. However, in most biological contexts, the assumption of independent features is certainly violated. For instance, genes regulated by the same set of transcription factors, or genes encoding enzymes for the same metabolic pathway, will tend to show correlated expression. Therefore, rather than treating each feature in an omics data set individually, it may be preferable to map from the high-dimensional molecular space to a much

smaller number of (possibly curated) functional biological networks. Mapping features into functional sets reduces dimensionality, increases the statistical power to detect small but coordinated disease perturbations, and improves the interpretability of the resulting molecular signatures.

In order to identify features that are associated with a clinical phenotype of interest, features can be mapped onto *a priori* defined and manually curated modules or "pathways". Gene Set Enrichment Analysis (GSEA) [91] is a very widely used approach to investigate pathway-level changes in gene expression data, and more recent proposals have also been made. One recently developed approach to identifying pathway-based molecular signatures for phenotype classification is the Differential Rank Conservation (DIRAC) method [92]. Unlike GSEA or other enrichment methods that usually return *p*-values for gene set enrichment, DIRAC builds a network-based molecular signature that identifies robust differences in pathway activity between two disease states.

However, one major caveat to such pathway-based approaches is that *a priori* defined pathways do not fully represent the complexity of the underlying biology, and may not be accurate within the particular physiological context. To overcome this limitation, molecular features can be mapped into more comprehensive interaction networks, such as protein-protein or protein-DNA interaction networks, which can be much more comprehensive and unbiased, as well as disease and context specific. Specifically, biological networks can be used as a structured framework to integrate omics data for the purpose of molecular signature development. For example, Chuang *et al.* integrated microarray gene expression data with protein-protein interaction networks to identify network-based prognostic biomarkers for breast cancer metastasis, and generated novel hypotheses regarding cancer progression [84]. The average sub-network activity, defined in this study as a function of expression levels of genes that compose the sub-network, was used to predict clinical outcome of breast cancer specimens. The network-based markers displayed better predictive accuracy on an independent dataset than markers selected without network information. In another study, Nibbe *et al*. used proteins that were differentially expressed between normal and cancer colon tissue from proteomics experiments as seeds to identify sub-networks enriched in these differentially expressed proteins from the human protein interaction network [93]. Then, the mRNA expression profiles of the components of these sub-networks were used as input features to a support vector machine in order to classify

colorectal cancer and normal samples. The prevalence of these networks being perturbed in colon cancer was demonstrated by these features alone being sufficient to achieve 90% classification accuracy in independent validations.

In the particular case of prion disease, a set of neurodegenerative disorders caused by the misfolding of prion proteins in the brain, Hwang *et al.* analyzed the dynamic network perturbations during the onset and progression of disease [86]. In this study, infectious prion proteins were delivered into the brains of living mice, and were harbored within the tissue for different time-spans of disease progression. At the end of each time-point, gene expression measurements were taken from harvested diseased brain tissue, and subsequently mapped onto physical protein interaction networks for comparative analysis. Intriguingly, this study showed reproducible perturbations that occurred in core networks that could be monitored prior to the manifestation of disease symptoms.

In the work summarized above, thousands of feature measurements for static biological states were used to characterize molecular networks. However, a more complete understanding of molecular networks requires perturbing the biological system under study in order to understand how the network components, as well as the clinical phenotype of interest, are affected by those perturbations. For example, stimulating one or more signaling pathways using *in vitro* cytokine assays can lead to different immunologic and metabolic responses in different diagnostic phenotypes [94], such as different disease progression levels. In a study by Hale *et al*. [95], the investigators used a cocktail of cytokines and mitogens to stimulate whole blood cells from patients with different stages of systemic lupus erythematosus, an autoimmune disease. They then used flow cytometry to measure multiple signaling responses at the single-cell level, generating a highly multiplexed view of intracellular signaling network activity during disease progression. They found that robust changes in signaling protein interactions in response to stimuli were good indicators of disease stage. Therefore, evaluating cell response after an activating stimulus may serve as a compelling approach for incorporating perturbations into patient classification going forward.

2.7 Are my features truly correct?

Given that two molecular signatures seem to perform well on independent data sets, how can we decide which is better? If all else is equal, we should prefer the molecular signature for which there is a plausible biological mechanism, as such a signature is much more likely to hold up in future patient samples as opposed to having overfit the data used in its development. Ideally, if sufficient numbers of samples were available, then a molecular signature's performance on one or many independent data sets would be the preferred way of assessing its suitability, regardless of whether or not a mechanism for its performance is known. But in reality, sample sizes are limited, and thus a molecular signature for which there is a plausible biological mechanism tends to be more convincing than one for which no such mechanism is known. Such biologically motivated signatures can also hold great promise to be developed as companion diagnostics for therapies, which may be motivated by the underlying mechanism. Thus, while lack of a known biological mechanism underlying a molecular signature certainly does not preclude its use provided that it works well in practice on independent samples, mechanistic information can increase our confidence that the signature will hold up to further scrutiny.

2.8 Pervasive bias in reported results

Another major challenge in omics-based molecular signature discovery is the prevalence of overly optimistic accuracies in reported results. This problem is not unique to omics research but is problematic in many data-driven research settings [96]. Such bias can occur for a number of reasons: 1) research groups tend to report only the best results among many attempted approaches; and 2) only positive results are published. Consequently, across the literature there is an overly optimistic view of how well molecular signatures perform. This pervasive bias is not necessarily the result of faulty science in any particular lab, but rather is a consequence of the way in which science is conducted and reported. This is responsible, in part, for the fact that many reported molecular signatures have not held up in follow-up studies.

2.9 The future of molecular signature discovery

In the future, we envision the development of molecular signatures using large publicly-available repositories of data, coupled with unbiased assessment of the successes and failures of these signatures. An automated system will integrate all available data for a clinical phenotype of interest, identify the most accurate molecular signature using a standard set of computational algorithms, and continuously update the signature as new data become available. The candidate molecular signature and all relevant performance results (e.g. overall accuracy, sensitivity, and specificity) will be reported and tracked over time. Once the molecular signature stabilizes, the system will eventually report a final molecular signature for the phenotype of interest. The results obtained from such an automated system will be unbiased, in the sense that both positive and negative outcomes (e.g. correct and incorrect predictions in the case of a categorical phenotype) will be recorded and reported. By integrating huge amounts of publicly available data, such a system will avoid some of the issues associated with batch effects and confounding that arise when smaller sample sizes are used for molecular signature discovery. Such a system will allow us to develop the most accurate possible molecular signatures and assess their performances as objectively and comprehensively as possible.

2.10 Conclusions

In this article, we have discussed some of the key considerations and challenges facing the discovery of omics-based molecular signatures of clinical phenotypes, such as good experimental design, careful data procurement, avoidance of over-fitting, validation on independent data sets, and integration of multiple data sets and data types. For guidance to the reader, Figure 2.4 summarizes the key steps in molecular signature discovery that were discussed throughout this article. We hope that this methodological checklist will aid investigators interested in identifying omics-based molecular signatures.

Since the emergence of the field of omics-based molecular signature discovery, researchers have developed an improved understanding of how to discover (and how not to discover!) such signatures. The field is still young, and as time passes, best practices in this area will continue to evolve. Currently, the number of validated and useful molecular signatures is disappointingly (but not surprisingly) small relative to the number of signatures that have been reported in the literature. However, we remain optimistic that as experimental and analytical practices improve, as sample sizes increase, and as techniques for data type integration continue to develop, omics-based molecular signatures will indeed transform the practice of medicine.

2.11 Chapter 2 figures

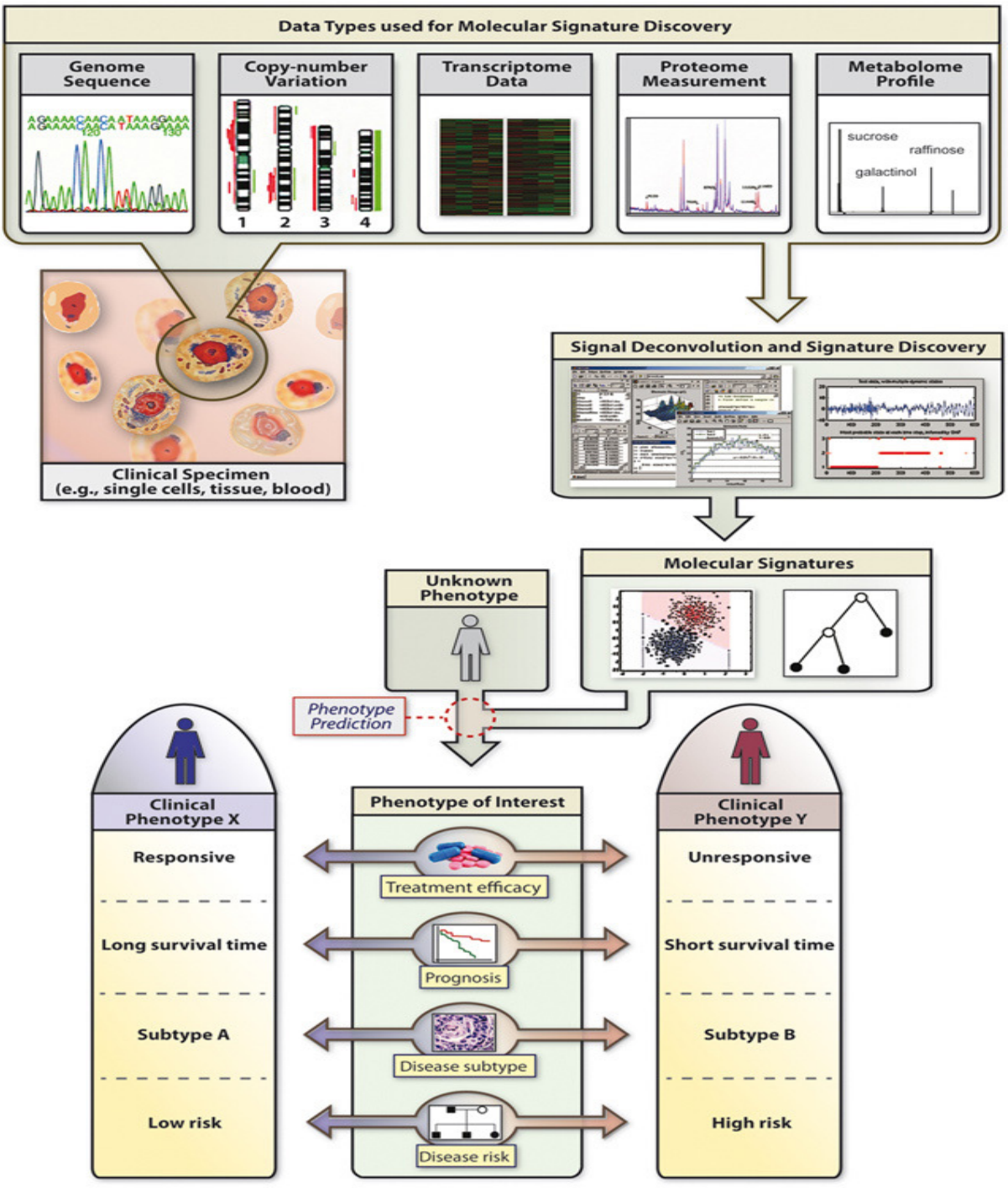

Figure 2.1 Overview of the discovery and application of molecular signatures from omics data. Molecular signatures can be derived from a broad range of omics data types (e.g. DNA sequence, mRNA, and protein expression) and can be used to predict various clinical phenotypes (e.g. response to therapy, prognosis) for previously unseen patient specimens.

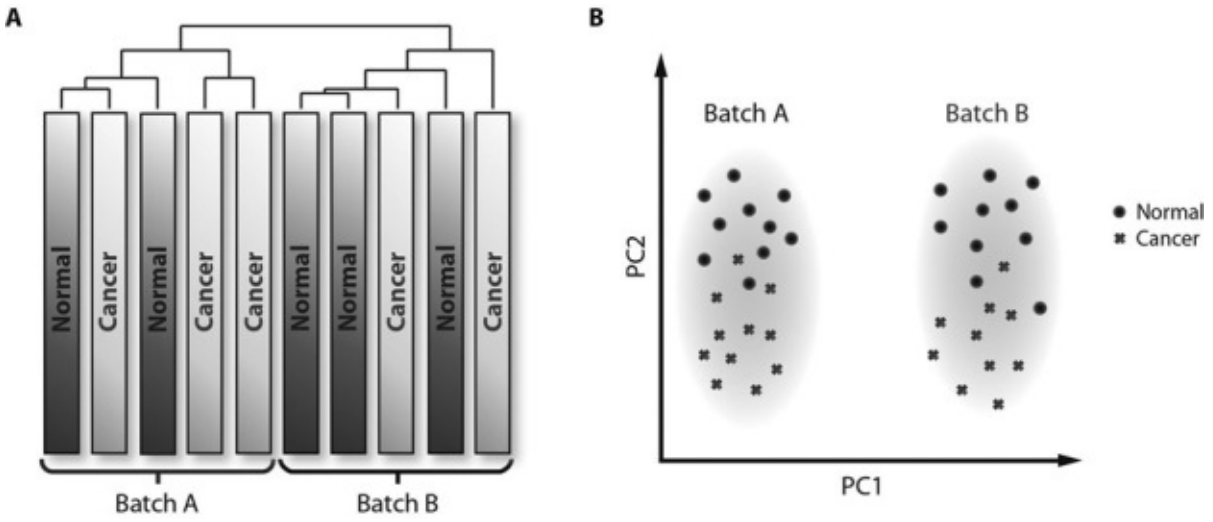

Figure 2.2. Two hypothetical scenarios in which (A) hierarchical clustering and (B) principal components analysis reveal that covariates other than the clinical outcome of interest have resulted in considerable discrepancies between patient populations. Here, batch characteristics and not group labels (cancer versus normal clinical specimens) are responsible for most of the observed variation among the samples. Such batch effects can arise due to changes in experimental protocols, data-processing techniques, or laboratory personnel at any point in the experimental process.

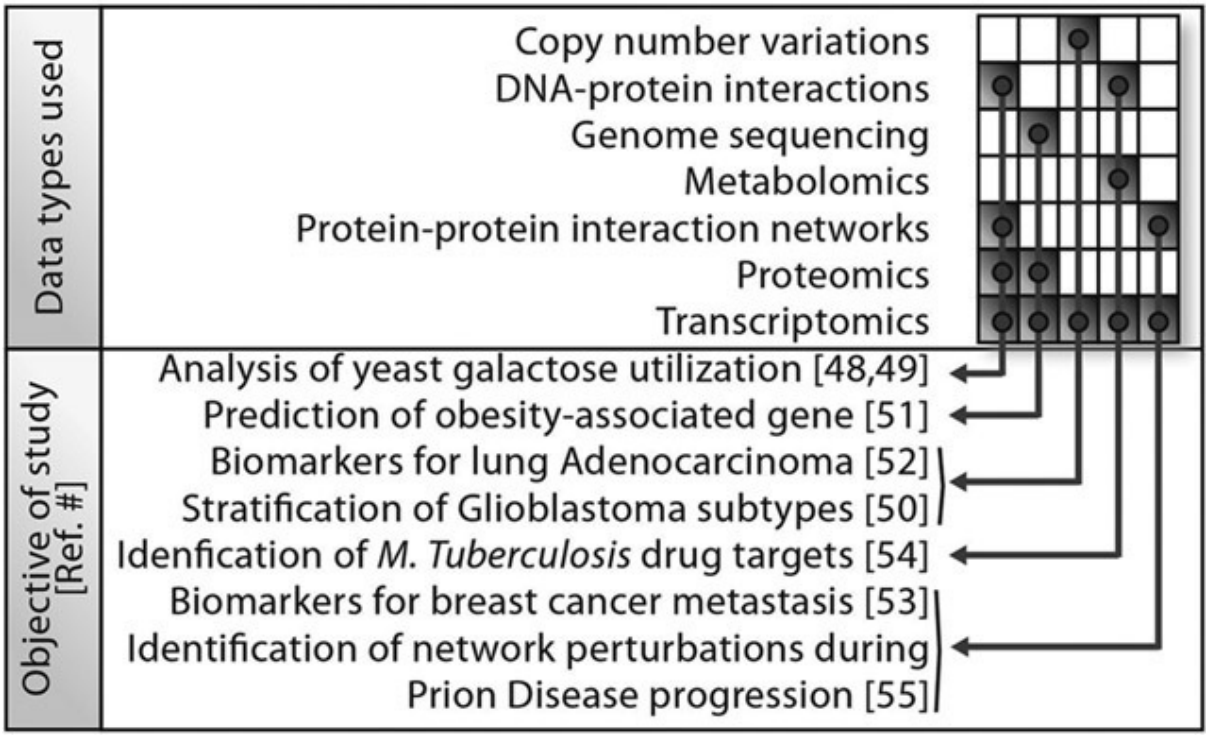

Figure 2.3 Combining different types of data across different measurement platforms can lead to more accurate molecular signatures for characterizing or predicting clinical phenotypes. Rows and columns of the checkered box correspond to data types and published studies, respectively. The collection of gray boxes in each column represents the combination of data types used in a particular study. The arrows designate the objective of each study.

Step 1. Establishing the scientific and clinical context

- Clearly define clinical phenotypes of interest
- Ensure that, if discovered, a molecular signature has the potential to be useful in the clinic
- Only use types of omics data that are suitable for addressing the task of interest
- Determine acceptable sensitivity and specificity

Step 2. Collecting omics data for molecular signature discovery

- When collecting new experimental data, ensure that:
  - sufficient sample size can be obtained
  - all aspects of the experimental and analytical procedures are carefully controlled to avoid batch effects
  - no confounding occurs between data sets of different phenotypes from factors unrelated to phenotype of interest
- When using existing data, ensure that:
  - sufficient sample size can be obtained
  - sufficient patient information is available for omics samples
  - proper normalization is implemented to make samples comparable across different data sets
- Consider integrating multiple data sets and data types:
  - approach with caution
  - can lead to molecular signatures that are more accurate and robust

Step 3. Developing molecular signatures through feature selection and model building

- Perform feature selection in either a supervised or an unsupervised manner
- Choose models that are well-suited for the context of the study and nature of phenotypes of interest
- Consider mapping features onto biological pathways or more comprehensive interaction networks
- Consider choosing models that show clear insight into plausible biological mechanisms
- Ensure that all cross-validation steps are performed correctly
- Approach favorable cross-validation results with caution

Step 4. Evaluating performance on independent datasets

- Test promising molecular signatures on independent data sets
- Independent test sets are not created equal. The strength of evidence from an independent test is based on the characteristics of the independent data set used (i.e. evaluating on data from multiple, different sites is a more stringent test than evaluating on data from only the same institution)

Step 5. Disclosing information on all aspect of study to enhance reproducibility

- Encourage the evaluation of the molecular signature by independent research groups
- Disclose: Information on the clinical context in which molecular signature is intended, patient selection criteria, clinical data (i.e. patient information), raw data, meta-data (if applicable), data processing and normalization methods, feature selection and model building methods, experimental protocols, records on study run-dates, lab technicians, reagent sources, etc., analytical methods, and source code

Step 6. Reporting all performance results to mitigate bias in public literature

- Encourage the objective assessment of molecular signatures by reporting both positive and negative outcomes (i.e. correct and incorrect predictions, respectively)
- Make data publicly available after publication

Figure 2.4 Steps for the development of molecular signatures on the basis of omics data.

Chapter 3. Relative mRNA levels of Functionally Interacting Proteins are Consistent Disease Molecular Signatures

In this chapter, I will describe a new computational tool, interacting Top Scoring Pairs (iTSP) that uses the relative mRNA expression reversal of two functionally interacting proteins between phenotypes as classifiers. This method extends a family of method that uses relative expression reversal of *k* gene pairs by constraining the gene pairs to be functionally related or physically interacting in a protein-protein interaction (PPI) network. I will first review existing Relative Expression Analysis (RXA) methods, and then present the iTSP method and comparison to its predecessor, the original Top Scoring Pairs (TSP) method.

3.1 Introduction to Relative Expression Analysis family of methods

Many machine learning methods have been applied to gene expression microarray data to identify molecular signatures that can accurately predict disease diagnosis, prognosis, and treatment response. Such methods often use complex decision rules with many tunable parameters that are not easily amenable to biological interpretation. The numerous possible combinations of data normalization methods and machine learning methods further complicate the choice of selecting the optimal combination. Relative Expression Analysis (RXA) methods uses the relative expression reversal of a gene pair between phenotypes (e.g., gene i expression is consistently higher than gene j in class 1, but consistently lower than gene j in class 2) as classifiers.

RXA methods have three advantages over other classification methods. First, as ranks of expression levels are used in place of expression values, RXA methods are invariant to the choice of all normalization methods that do not scramble these ranks. Second, RXA methods use only a small number of features and requires little to no parameter tuning. This may facilitate the development of inexpensive clinical tests using RT-PCR that only need to measure expression level of a few genes. Third, the decision rule is very straightforward to interpret.

The foundation of all RXA methods is the original TSP method [97]. Let $P_{ij}(C_m)=Pr(X_i>X_j|Y=C_m)$ be the probability of observing $X_i>X_j$ (the expression level of gene i is higher than that of gene j) in class C_m. $P_{ij}(C_m)$ is estimated by the frequency of observing $X_i>X_j$ in class C_m. Let $\Delta_{ij}= |P_{ij}(C_1) - P_{ij}(C_2)|$ be the score of each gene pair (i, j), which quantifies the difference in probability of observing $X_i>X_j$ between class 1 and class 2. A score of 0 means that the ordering $X_i>X_j$ is equally likely in both classes, and the relative ordering of gene pair (i, j) is not informative of class distinction; a score of 1 means we always observe $X_i>X_j$ in class 1 and never in class 2 (or vice versa), and the relative ordering of gene pair (i, j) is highly informative of class distinction. The higher the score, the better the gene pair (i, j) can classify class 1 and 2. The original TSP method identifies the gene pair that achieves largest Δ_{ij} as classifier. Although the combinatorial search space is very large, few randomly selected gene pairs achieve similar scores as the true TSP [97].

k-TSP extends TSP by using k disjoint gene pairs, where k is a tunable parameter chosen by cross-validation (when k=1, k-TSP is reduced to original TSP) [98]. Decision is based on majority voting among k gene pairs. Tan *et al* compared k-TSP and TSP to commonly used machine learning methods such as k-Nearest Neighbor (kNN), Support Vector Machines (SVM), Decision Trees (DT), Naïve Bayes (NB), and Prediction Analysis of Microarrays (PAM) in 19 different human cancer data sets [98]. k-TSP and TSP perform similar as PAM and SVM. These four methods perform better than the remaining methods. Besides similar performance, k-TSP and TSP has the advantage of simple decision rules and small number of features (at most 2k) used. Top Scoring Triplet (TST) extends TSP by using relative expression reversals among *triplet* of genes as classifiers [99]. Top Scoring 'N' (TSN) extends TSP and TST by using relative expression reversals among n genes (n=2, TSP; n=3, TST, n=4, top scoring quadruple, etc) [100].

3.2 interacting Top Scoring Pairs: Putting Biological Constraints on Top Scoring Pairs

Despite its high predictive performance and simplicity, TSP has one major drawback: when it is applied to two independent data sets with the same classification endpoint, there is little or no overlap in the top 50- or 100-gene pair lists. Low reproducibility in classification methods is not a problem specific to TSP. Different partitioning of the same dataset can result in different sets of marker genes

for many classification algorithms [101], and the consistency of marker genes across independent data sets is even lower [102]. In this chapter, I will describe a method called *interacting Top Scoring Pairs* (iTSP), which integrates functional protein interaction networks (e.g., interactions in the STRING database [103]) with transcriptomic data to identify high confidence interacting proteins for which the *relative* expression of the corresponding genes is consistently reversed between phenotypes. iTSP retains the desirable features of TSP, including straightforward decision rules, independence of most normalization procedures, and a prediction rule based on only two genes. iTSP also achieves a predictive performance that is comparable to that of TSP. However, through the integration of functional protein interaction networks, iTSP dramatically increases the consistency of gene pair selection in comparison to TSP. Moreover, unlike previous network-based classification methods, iTSP also accounts for interaction quality and automatically selects high-quality interactions as classifiers. Finally, iTSP identifies function protein interactions that can readily serve as a basis for generating novel hypotheses about disease processes.

Both TSP and iTSP use a score Δ_{ij} that quantifies the separability of two phenotypes C_1 and C_2 based on the relative expression level of two genes *i* and *j* [97]; see Methods. The score assumes values between 0 and 1 and measures how informative the order of expression of the two genes is about the true phenotype: $\Delta_{ij} = 0$ corresponds to no information while $\Delta_{ij} = 1$ indicates perfect discrimination. The classifier iTSP differs from TSP in two ways. First, iTSP only considers functionally interacting gene pairs. Second, in gene pair identification, instead of maximizing Δ_{ij} alone, iTSP also takes into account a confidence score *S* for the interaction between the gene pair:

$$\Delta'_{ij} = (1 - \alpha) * \Delta_{ij} + \alpha * S,$$

where α is an adjustable parameter. In the STRING database, the confidence score, S, of interactions ranges from 0.15 to 1, which corresponds to the probability of finding the linked proteins within the same KEGG pathway [103]. A larger α biases the search for a gene pair towards higher confidence interactions, and the parameter is selected with internal cross-validation. In this study, around 400,000 interactions from the STRING database were considered. To ensure a fair comparison of iTSP and TSP, the total number of possible gene pairs for each method was kept the same.

3.3 iTSP is comparable in predictive performance to TSP

We evaluated the predictive performance of iTSP in two ways. We first did 10 repeats of 10-fold cross-validation for each of 6 binary classification endpoints on a liver data set consisting of 4 liver phenotypes: normal liver (39 samples), chronic hepatitis C (CHC, 36 samples), cirrhosis (CH, 143 samples), and hepatocellular carcinoma (HCC, 171 samples). Predictive performance was measured by the Matthews Correlation Coefficient (MCC), as recommended by the MAQC-II study; MCC values range from +1 to −1, with +1 indicating perfect prediction, 0 indicating random prediction and −1 indicating perfect inverse prediction[104]. As a control, we performed permutation testing that kept the graph fixed, i.e., the number of edges impinging on each node in the PPI unchanged, but shuffled the protein (node) labels to generate a random network with equivalent topological properties. Thus, each resulting random network had exactly the same node degree distribution as the real PPI. The randomized PPI networks were used in iTSP. The average MCCs of iTSP, iTSP with randomized PPI, and TSP across 6 binary classification endpoints were 0.755±0.017, 0.724±0.034 and 0.773±0.022, respectively (Figure 3.1 A).

As a second and more stringent test of predictive performance, we also did independent validation of iTSP and TSP on 6 MAQC classification endpoints, including a positive control (endpoint H) and a negative control (endpoint I). For each endpoint, there was an independent training and validation data set. We trained on the training data set, evaluated predictive performance on the validation data set, and recorded MCC. We then swapped the training and validation data sets and repeated the above process. The average MCC of each endpoint is shown in Figure 3.1B. We also plotted the median of the 17 best MCCs recorded for each endpoint ("Median of MAQC"), which is used in the MAQC-II study as a measure of the inherent predictability of each endpoint[104]. iTSP and TSP had similar predictive performance across 6 classification endpoints and the predictive performance depended on the inherent predictability of the endpoints. The mean MCCs of iTSP, iTSP with randomized PPI, TSP, and "median MAQC" across 5 binary classification endpoints (negative control I is excluded) were 0.431, 0.395, 0.412, and 0.441, respectively The MCC for iTSP with randomized PPI was the average from 1000 randomized PPI networks. These results show that iTSP performs comparably in classification to both TSP and the MAQC-II algorithms.

Therefore, in both cross-validation and independent validation, iTSP had comparable predictive performance to TSP.

3.4 iTSP significantly increases transcriptomic signature selection consistency

In addition to comparable predictive performance, iTSP significantly increased gene pair selection consistency. We used two approaches to assess gene signature consistency. The first approach was cross-validation. It has been reported [101] that the particular gene signatures identified within each iteration of cross-validation are strongly dependent on data splitting, because i) many genes are correlated with the endpoint, ii) the differences in correlation are small and iii) correlation fluctuates as different subsets are chosen for training and testing. To evaluate the ability of iTSP to yield consistent gene signatures across different training-testing data splitting, we did 10 repeats of 10-fold cross-validation. To quantify the consistency of gene pair selection, we calculated the fraction of times the same TSP was chosen in two loops within 10-fold cross-validation over the 45 possible pairs of loops ($C_{10}^2 = 45$), and averaged the probability across 10 repeats of 10-fold cross-validation. **Figure 3.2** shows that iTSP had higher reproducibility than TSP in 5 of the 6 liver classification endpoints. However, iTSP with real PPI did not always generate more consistent gene pairs than iTSP with randomized PPI, which is not surprising since severely limiting the set of gene pairs to choose from, whether real or randomized, makes iTSP less susceptible to fluctuations in training data splits and score estimation. As iTSP automatically favors high-quality interactions, the median STRING score of the 600 interacting gene pairs selected by iTSP (over 10 runs of 10-fold cross-validation for 6 classification endpoints) is dramtically higher than the median score of all input protein-protein interactions to iTSP (0.90 v.s. 0.66, Wilcoxon ranksum test p-value 7.8×10^{-102}). Therefore, the benefit of using real functional interactions is that the selected gene pairs are highly likely to participate in the same biological pathways, and may point to specific hypotheses about disease perturbation.

The second approach for assessing reproducibility of gene signatures was to use independent data sets for the same classification endpoints. In the MAQC study, for the same endpoint, two independent data sets (training and validation) were generated from different patient cohorts[104]. We applied iTSP and TSP to each independent dataset and compared the overlap of gene signatures learned from

the different datasets. Table 3.1 shows the overlap of gene pairs for two independent data sets of endpoints D, E, and H using iTSP and TSP, respectively. If there was a strong signal in the microarray data and the endpoint was easy to classify, iTSP was dramatically more consistent (i.e., more overlapping gene pairs) than TSP. This is the case for endpoint E and H. For endpoint H, which was the positive control (sex of multiple myeloma patients), 44 out of the top 50 gene pairs identified by iTSP from two independent data sets were the same, while there was no overlapping gene pair identified by TSP. When the signal was weaker and the endpoint was harder to classify, such as in endpoint D (success of treatment involving chemotherapy followed by surgical resection of a breast tumor), the consistency of iTSP decreased, but it was still much higher than that of TSP.

Throughout the analysis, functional protein interactions from the STRING database are used, and iTSP automatically favors interactions with high STRING quality scores, and therefore high probability of being in the same pathway. It is possible that iTSP achieves higher consistency by selecting more redundant gene pairs (i.e. genes that are highly correlated with each other). We compared the Pearson correlations of 34 unique gene pairs selected by iTSP in 10 repeats of 10 fold cross-validation across 6 liver classification problems with 64 unique gene pairs selected by TSP. The correlation is calculated separately for each classification problem (e.g., expression correlation of gene pairs selected to classify HCC v.s. normal was calculated based on HCC and normal microarray samples). The difference in correlation is not statistically significant: median absolute Pearson correlation of 34 iTSP pairs is 0.235, while median absolute Pearson correlation of 64 TSP pairs is 0.325 (Wilcoxon ranksum test p-value: 0.217). The same results hold if Spearman correlation is used. Therefore, even though iTSP favored gene pairs with higher STRING confidence scores and therefore higher probability of in the same pathway, the resulting features are no more statistically redundant (i.e. correlated) than choosing gene pairs without any functional constraint, as done in TSP.

3.5 Favoring high-quality interactions increases transcriptomic signature accuracy and consistency

In addition to considering the level of discrimination Δ_{ij}, iTSP also accounts for the confidence score S of the interaction between the gene pair (i, j). This joint optimization criterion can be viewed as a form of regularization that balances predictive performance with biological relevance of the gene pair. To evaluate the effects of favoring high-quality interactions, we set the regularization parameter to zero and selected gene pairs purely based on the original Δ_{ij} score. The average MCCs of iTSP and iTSP without regularization using 10 repeats of 10-fold cross-validation across 6 binary classification endpoints on the liver data set were 0.755 and 0.715, respectively (Wilcoxon ranksum test p-value 3.3×10^{-4}, Figure 3.3A). In independent validation on the MAQC data set, the average MCCs of iTSP and iTSP without regularization across 5 binary classification endpoints were 0.431 and 0.406, respectively (Figure 3.3B). Removing regularization also dramatically reduced gene pair selection consistency in 10 repeats of 10-fold cross-validation on the liver data set (Figure 3.4). Therefore, we showed that favoring high-confidence interactions in iTSP improves both predictive performance and the consistency of transcriptomic signature selection. To our knowledge, iTSP is the first classification method to explicitly account for the quality of protein interactions, which is important because both functional and physical protein interaction networks include many false positive interactions.

3.6 iTSP identifies protein-protein interactions related to disease processes

As the protein products of gene pairs picked by iTSP functionally interact with each other, we can use iTSP results to propose interesting hypotheses about perturbed protein-protein interactions related to disease processes. For example, the CYP2C19-GSTP1 gene pair is an accurate classifier between chronic hepatitis C and cirrhosis. Both genes are involved in xenobiotic metabolism, and their interaction has a confidence score of 0.95 in the STRING database, which is at the highest confidence level. CYP2C19 activity is known to be modulated by cirrhosis [105]. The poor metabolizer phenotype of the CYP2C19 genotype is more frequent HCV-related cirrhosis [106]. Polymorphisms in GSTP1 are known to affect alcoholic [107] and cryptogenic cirrhosis[108]. The relative expression reversal of CYP2C19 and GSTP1 (Figure 3.5) may indicate aberrations in xenobiotic metabolism as chronic hepatitis C progresses to cirrhosis.

As a comparison, the TSPAN9-RHOG gene pair was often selected by TSP as an accurate classifier of CHC v.s. CH. While CYP2C19 and GSTP1 shared two Gene Ontology (GO) annotations, GO:0044281 (small molecular metabolism, containing 1300 genes) and GO:0006805 (xenobiotic metabolism, containing 140 genes), TSPAN9 and RHOG only shared one very generic GO annotation, GO:0016020 (membrane, containing 2240 genes). To systematically quantify whether iTSP-selected gene pairs share more specific GO annotations and potentially lead to more specific biological hypotheses, we used the Fisher's Omnibus statistic (Methods, and [109]). A large Fisher's Omnibus statistic means that two genes share a large number of specific GO terms. The median Fisher's Omnibus statistic of 600 gene pairs generated by for iTSP and TSP (including redundant gene pairs) in 10 repeats of 10-fold cross-validation for 6 liver classification problems were 6.65 and 1.37, respectively (Wilcoxon ranksum test p-value 2.2×10^{-53}). As Gene Ontology is among the many information sources to calculate interaction confidence scores in the STRING database, the higher functional coherence of iTSP-selected gene pairs is partially because of its bias toward high quality interactions. The above analysis is therefore not an independent validation of the biological significance of iTSP-selected gene pairs, but a demonstration that incorporating functional information in classification returns more biologically interesting classifiers than just choosing the most accurate classifier.

3.7 Conclusion

We developed the *interacting Top Scoring Pairs* (iTSP) method, which integrates protein interaction networks with transcriptomic data to identify functionally interacting proteins for which relative expression levels are reversed between phenotypes. While achieving predictive performance comparable to TSP, iTSP yields more consistent and biologically meaningful gene signatures.

While previous methods of integrating protein interaction network with transcriptomics data typically have focused on identifying sets of nodes for which expression level *per se* changes between classes [110], fewer methods have focused on identifying perturbed protein *interactions*. iTSP is a novel method that identifies functional interactions that are accurate phenotype classifiers and indicative of interesting biological processes related to disease. Protein interaction networks are known to include many false positive interactions. iTSP improves upon previous network-based classification methods by explicitly favoring high-quality interactions with a regularization parameter. This approach is applicable to both functional protein interactions in the STRING database or physical protein interaction network in the HIPPIE database[111], where quality scores are available.

3.8 Method

Detailed description of iTSP

iTSP scores each functionally interacting gene pair by a combination of Δ_{ij} and confidence score of the interaction: $\Delta'_{ij} = (1-\alpha) * \Delta_{ij} + \alpha * S$, where S is the interaction confidence score and α is the regularization parameter. Optimal α is determined by 5-fold cross-validation within the first iteration of 10-fold cross-validation and used for all 10 iterations. Genes for which expression level were below 100 in more than 90% of samples (ignoring class labels) were filtered inside cross-validation on training samples before iTSP or TSP were applied. iTSP will be included in future release of Adaptive Unified Relative Expression Analyzer (AUREA)[112].

Fisher's omnibus procedure

Fisher's Omnibus statistic was used to quantify the number and specificity of shared GO annotation terms between two genes. A large Fisher's Omnibus statistic means that two genes share a large number of specific GO terms.

$$F_{ij} = \sum_{g_i, g_j \in T_m} -2log\,(\frac{M_t}{N})$$

T_m is a GO annotation term. M_t is the number of genes in T_m, and N is the total number of genes with GO annotations. Therefore, genes that share *many specific* terms will have high similarity scores and tend to participate in the same biological processes.

Description of data sets used in this study

The liver data set. We collected 392 transcriptomes from 7 different studies [113-119] from 6 distinct labs. This data set consisted of normal liver, chronic hepatitis C (CHC), cirrhotic liver and HCC. We downloaded the publicly available raw intensity files of these 392 samples from Gene Expression Omnibus and used our own consensus preprocessing pipeline to process the raw data. This pipeline went through all three different Affymetrix platforms (U133A, U133 2.0 and U133 plus 2.0), found common probe sets across platforms, built a consensus platform, and used the Matlab implementation of GCRMA [120] to preprocess all samples together. After this preprocessing step, we had 22,277 probes in common among the three Affymetrix platforms. Probes that did not map to any known genes according to the latest annotation were removed. After this step, we had 20,928 probes. Probes were mapped to Entrez gene IDs according to the latest Affymetrix annotations, and if there were multiple probes mapping to the same gene, the maximum expression value was used. After this step, there were 12725 Entrez gene IDs. ComBat[121] was used to mitigate batch effects arising from combining data sets of different sources.

The MAQC data set. The preprocessed MAQC data set was downloaded from[122], and the mapping of probe to gene was done as described for the liver data set.

The functional protein-protein interaction network. The human functional protein interaction network was extracted from the STRING database version 9.05. Interactions with confidence score above 0.4

(STRING database's threshold for medium confidence level) were used. Ensemble protein IDs were mapped to Entrez Gene IDs using Ensemble Biomart. Redundant interactions at the gene level were removed, resulting in 401,814 unique functional interactions between 16,560 genes.

3.9 Chapter 3 figures and tables

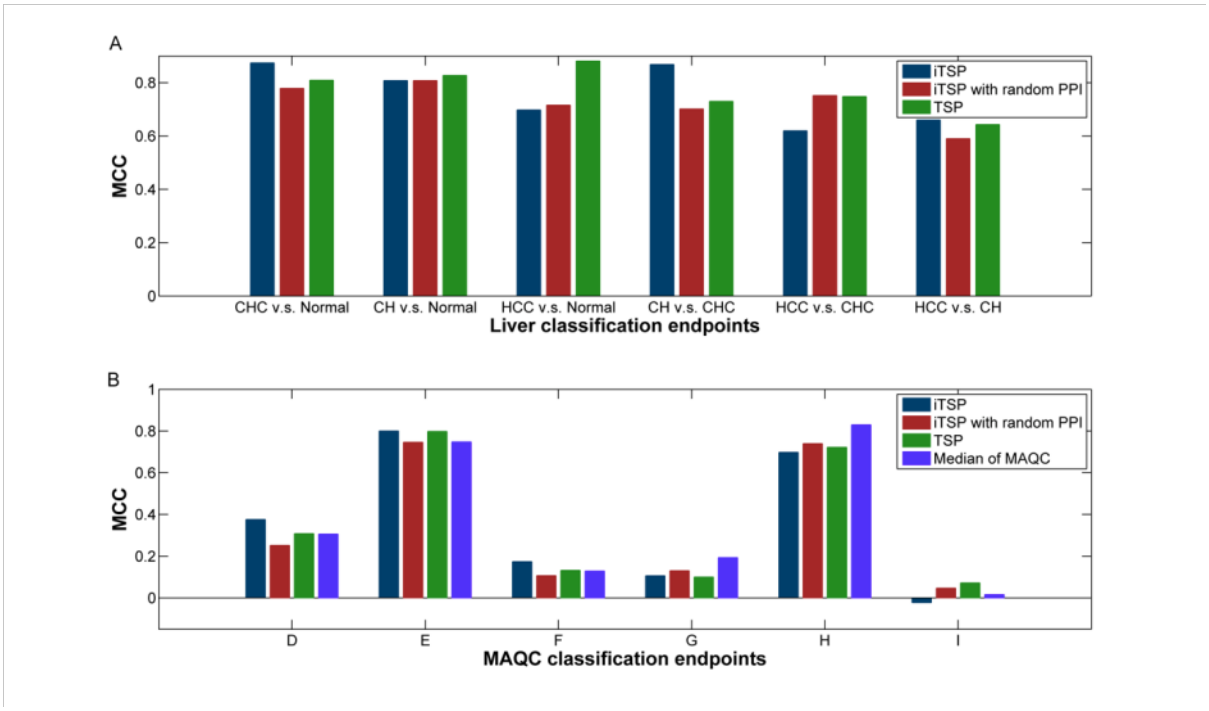

Figure 3.1 Comparison of predictive performance of iTSP and TSP. A. Predictive performance with a liver data set in 10 repeats of 10-fold cross-validation. B. Predictive performance with matched MAQC data sets (trained on one set, tested on the other).

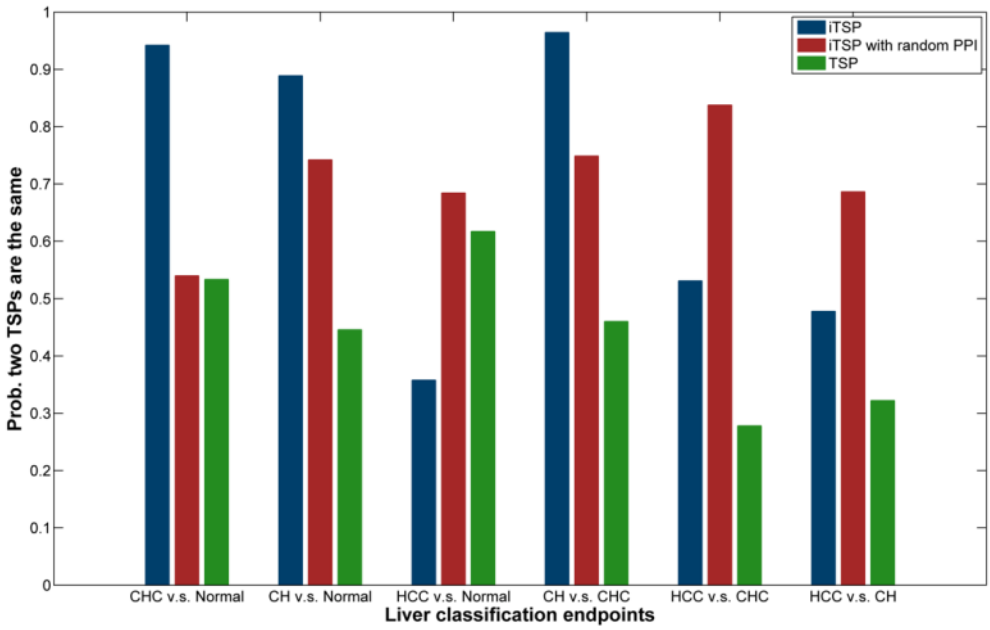

Figure 3.2 Comparison of gene pair selection consistency between iTSP and TSP. Consistency is measured by the probability that the pair of genes selected in two loops of cross-validation are the same.

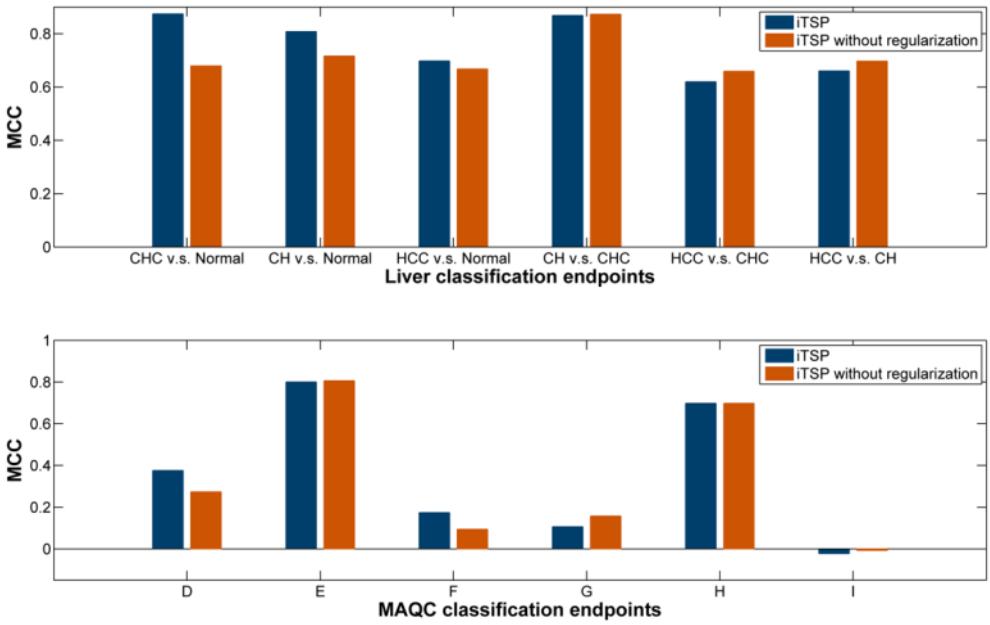

Figure 3.3 Comparison of predictive performance of iTSP and iTSP without regularization. A. Predictive performance on a liver data set in 10 repeats of 10-fold cross-validation. B. Predictive performance on MAQC data sets.

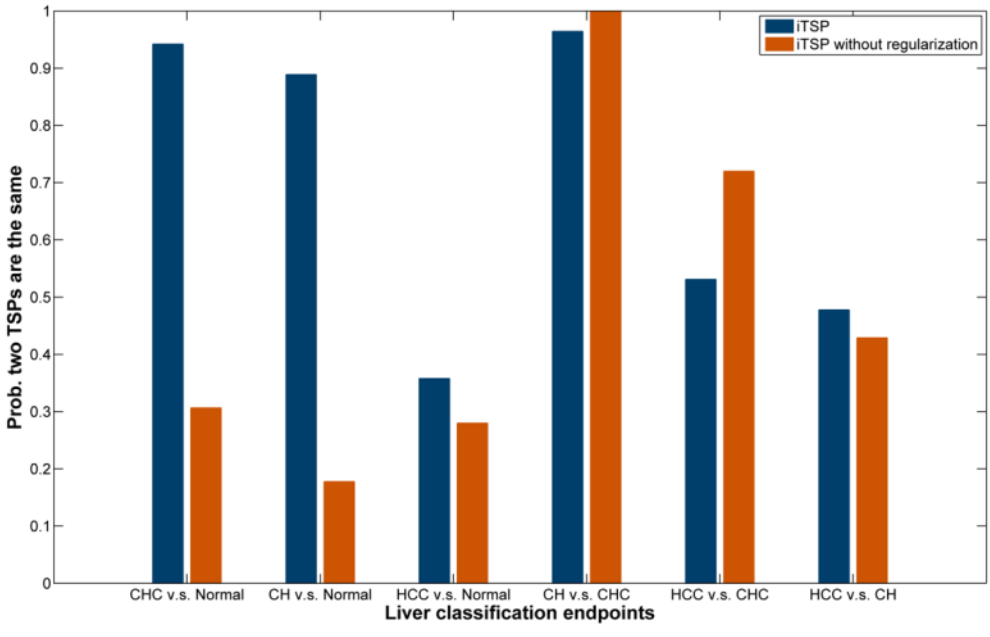

Figure 3.4. Comparison of gene pair selection consistency between iTSP and iTSP without regularization. Consistency was measured by the probability that two gene pairs selected by each method were the same within 10-fold cross-validation.

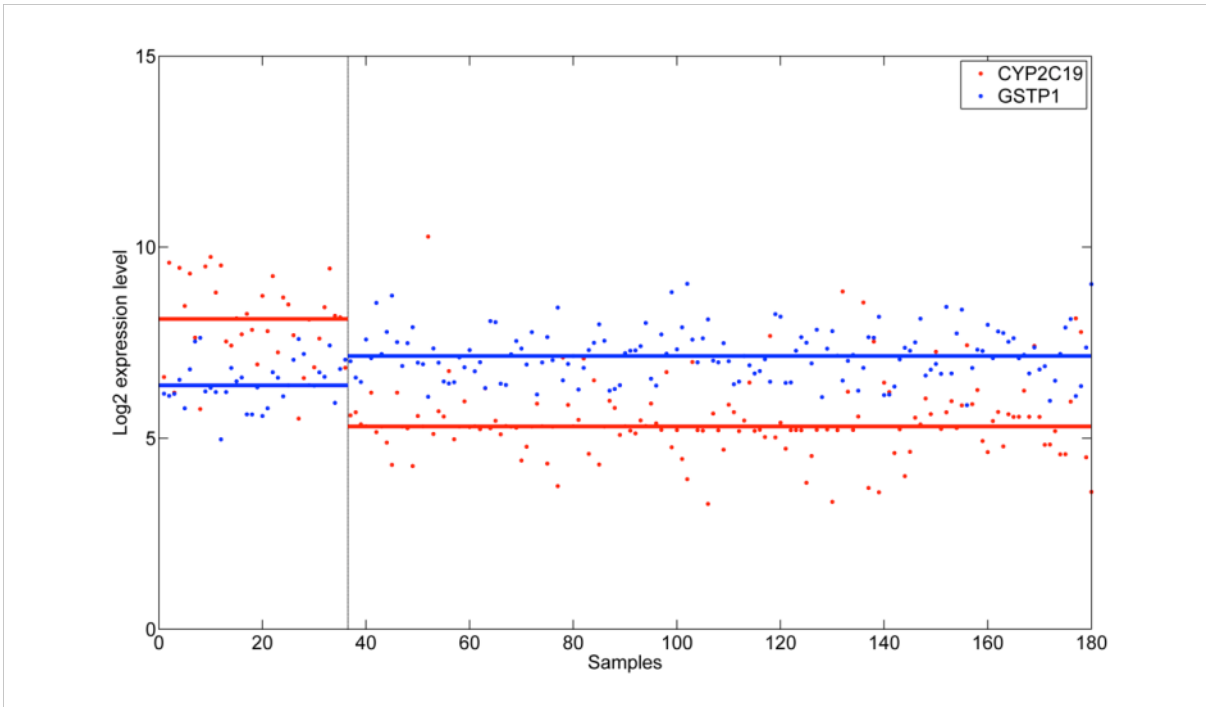

Figure 3.5. Expression of the GYP2C19-GSTP1 gene pair that accurately classifies chronic hepatitis C (CHC, left) and cirrhosis (CH, right). Horizontal lines represent median expression levels of genes within a class.

	Number of overlap gene pairs among top					
	Endpoint D		**Endpoint E**		**Endpoint H**	
	iTSP	**TSP**	**iTSP**	**TSP**	**iTSP**	**TSP**
Top 10	0	0	1	0	5	0
Top 20	0	0	2	0	17	0
Top 50	5	0	18	1	44	0
Top 100	15	0	41	1	48	0

Table 3.1. Consistency of gene pair selection measured by the number of overlapping gene pairs when each method was applied to two independent data sets of the same MAQC classification endpoint.

Chapter 4. Transcriptional shifts at metabolic branch points reflect phenotypic differences

4.1 Identify coordinated transcriptional changes in metabolic network

Metabolism is under extensive transcriptional regulation [200]. Cellular response to genetic and environmental perturbations often involves changes in metabolic activities, some of which in the form of differential expression of metabolic genes. For example, a recent comprehensive transcriptomic comparison of 22 cancer tissues and their cognate normal tissues revealed both common and tissue-specific gene expression changes in tumorigenesis [30]. Although gene set approaches such as the Gene Set Enrichment Analysis (GSEA) [18] can reveal coordinated transcriptional changes in a priori defined gene sets, it cannot account for the topological feature of metabolism: different pathways are interconnected. Reporter metabolite analysis [201]. This method defines a metabolite-reaction bipartite graph where a metabolite is linked to all reactions that it participates in (consumed or produced). Reporter metabolites are metabolites whose associated reactions are more significantly differentially expressed than expected by chance, and represent "hot spots" where transcriptional regulations occur. This method has been used to identify metabolites affected by genetic perturbations in microbes [201] as well as diabetes [202].

However, reporter metabolite analysis does not consider the directionality of reactions-whether a reaction consumes or produces the metabolite. Moreover, it only considers transcriptomic changes of reactions directly linked to a metabolite, while it is known that distant reactions can also have an impact. Therefore, I developed a new analysis pipeline that considers transcriptional changes at metabolic branch points. In particular, the analysis is currently restricted to reactions that *consume* the same metabolite. An example metabolic branch point is the cholesterol 25-hydroxylase and CYP46A1 reaction pair. At this branch point, cholesterol can either be converted to 25-hydroxycholesterol by cholesterol 25-hydroxylase or to 24-hydroxycholesterol by CYP46A1.

4.2 Transcriptional shifts at metabolic branch point reaction pairs can distinguish cancer and normal tissues

The first question I want to address is, can we distinguish cancer and normal tissue transcriptomes using expression profile at metabolic branch points. In this analysis, I used the relative expression level of two reactions that consume the same metabolite as classifiers. This is essentially an interacting Top Scoring Pairs (iTSP, Chapter 3) approach, where instead of gene pairs, we use reaction pairs, and instead of restricting to protein interactions, we restrict to reactions that consume the same metabolite. Figure 6.1 showed predictive performance (measured by Mathews Correlation Coefficient) across 13 different types of tumor v.s. normal classifications. The average MCC is 0.717 and accuracy is 0.875.

Table 4.1 listed branch point reaction pairs that show consistent expression shifts between cancer and normal tissues across many different types of tissues. Positive score denotes that $R_1>R_2$ (i.e., enzyme catalyzing reaction 1 is expressed higher than enzyme catalyzing reaction 2) is more frequent in cancer than in normal; negative score denotes that $R_1>R_2$ is more frequent in normal than in cancer. Reaction pairs are sorted by the sum of absolute scores across 13 tumor v.s. normal comparisons.

Catalase (CATm) and glutathione peroxidase (GTHPm) are two different anti-oxidant enzymes that convert hydrogen peroxide to water and oxygen. Table 6.1 showed that glutathione peroxidase is consistently expressed higher than catalase in tumor tissues, while catalase is consistently expressed higher than glutathione peroxidase in normal tissues. Previous experiments suggested that catalase is the primary anti-oxidant enzyme when intracellular hydrogen peroxide is low, while glutathione peroxidase is preferentially used when hydrogen peroxide concentration is high [203]. It is possible that as tumor tissues are subject to higher oxidative stress [204], glutathione peroxidase is preferred over catalase. In breast cancer MCF7 cell line, over expression of catalase results in reduced expression of glutathione peroxidase and cell proliferation, suggesting that the glutathione peroxidase is functionally important in breast cancer [205].

4.3 Metabolic heterogeneity at branch points

As discussed in Chapter 1, there is considerable heterogeneity in cancer metabolic profiles. Therefore, I analyzed the heterogeneity of reaction expression reversal at metabolic branch points. For each reaction pair, an entropy measure is defined as:

$$H= -\log_2(p_1) - \log_2(p_2),$$

where p_1 is the frequency of $R_1>R_2$, and p_2 is the frequency that $R_2>R_1$. Entropy is maximized when the relative expression level of the two reactions are randome: $R_1>R_2$ in 50% samples, and $R_2>R_1$ in 50% samples. Table 4.2 listed branch point reaction pairs that have higher entropy in tumor tissue samples than in normal tissue samples: while there is a consistent trend of $R_1>R_2$ (or $R_2>R_1$) in normal tissues, such consistent trend does not exist in the corresponding tumor tissues. This suggests that different subsets of tumors have different preference of R_1 v.s. R_2.

The reaction pair with the greatest entropy difference in tumor v.s. normal across most tissue types is GLGNS1 (glycogen synthase) and UDPGP (UDP glucose pyrophosphohydrolase). A recent study found that hypoxia induced an early increase in glycogen synthase expression, and glycogen utilization is essential to prevent premature senescence [206].

Another interesting branch point with high entropy in tumor is GTHS (glutathione synthase) v.s. PRAGSr （phosphoribosylglycinamide synthase. GTHS is a key step in the biosynthesis of glutathione, a key co-factor used by the anti-oxidant enzyme glutathione peroxidase. PRAGSr is a key step in inosine-monophosphate (IMP) synthesis, a component of nucleotides. Therefore, this branch point represents a balance between coping with oxidative stress and increase proliferation. Recent experiments identified that p53 and p21 are key regulators of flux split at GTHS and PRAGSr [207]. Under nutrient stress, cancer cells with wild type p53 and p21 can divert more flux via GTHS to combat increased oxidative stress, and grow significantly *faster* than cancer cells with mutant p53 or p21, which continue to divert more flux toward IMP synthesis, despite that IMP is a key biomass precursor. Higher entropy at this branch point suggests that a subset of tumor samples is limited by oxidative stress, while another subset is limited by biomass precursors.

The other side of the analysis is branch points with high entropy in normal tissue samples but low entropy in tumor tissue samples. These branch points may suggest that compared to normal tissues, tumor tissues are "addicted" to a particular reaction at the branch point and therefore show consistently higher expression in a particular reaction. Table 4.3 list such branch point reaction pairs.

The top reaction pair with low entropy in tumor tissues and high entropy in normal tissues is CYSTA (cysteine transaminase) v.s. PPNCL3 (phosphopantothenate-cysteine ligase). Different types of tumor tissues all consistently have higher expression level of PPNCL3 than CYSTA. PPNCL3 is a key step in Co-enzyme A (CoA) synthesis from pantothenate. As CoA is involved in many biosynthetic reactions, this result suggest that tumor tissues consistently favor PPNCL3 over CYSTA because CoA for proliferation.

Another reaction pair with low entropy in tumor tissues and high entropy in normal tissues is ACITL (ATP citrate lyase) and ACONT (aconitase). Different types of tumor tissues all consistently have higher expression level of ACITL than ACONT. ATP citrate lyase is a key step in de novo fatty acid synthesis and is reported to be over-expressed in many tumor types. Targeting ACITL has been shown to inhibit tumor growth [208], demonstrating the functional implication of this branch point expression "addiction".

4.4 Incorporating global constraints to identify preferred reactions at branch points

Relative metabolic flux split at a metabolic branch point needs not to be determined solely by the relative expression level of the corresponding enzymes, as global constraints, especially expression levels of upstream and downstream reactions can also have an impact. Therefore, to go beyond just using metabolic network topology, I also applied metabolic network functional constraints on flux split at metabolic branch points.

1. For each metabolic branch point reaction pair (R_1, R_2), find the maximum flux through R_1 and R_2 without any constraints other than mass balance and thermodynamic.

2. Maximize the agreement between metabolic flux and reaction expression, while forcing reaction R_1 to carry 90% of its maximum flux. The optimal fit is F_1, which is the total number of highly expressed reactions that carry flux and the lowly expressed reactions that do not carry flux.

3. Repeat step 2 for reaction R_2, get the optimal fit F_2

4. If $F_1>F_2$, then R_1 is the preferred reaction at this metabolic branch point; if $F_1<F_2$, R_2 is the preferred reaction; if $F_1==F_2$, there is no preference.

This analysis is repeated for each reaction pair in each sample of tumor and normal tissues. Each reaction pair will have a score:

$$\Delta_r = \Pr(F_1 > F_2 | Tumor) - \Pr(F_1 > F_2 | Normal)$$

If Δ_r is close to 1, then R_1 is consistently preferred over R_2 in tumor tissues; if Δ_r is close to -1, R1 is consistently favored over R_2 in normal tissues.

Table 4.4 showed the top 20 reaction pairs with the largest absolute Δ_r in pancreatic cancer. The top reaction pair adenine phosphoribosyltransferase (ADPT) and nicotinate-nucleotide diphosphorylase (NNDPR) is particularly interesting. ADPT is an important reaction in purine nucleotide salvage pathway that uses adenine and phosphoribosyl pyrophosphate (PRPP) to form AMP. NNDPR catalyzes NAD synthesis from its precursor quinolinic acid. NAD plays a vital role in cancer metabolism[209]. Recent experiments found that NNDPR is induced with oxidative stress and is associated with poor prognosis in glioma[210]. Another interesting reaction pair is adenylosuccinate synthase (ADSS) and argininosuccinate synthase (ARGSS). ADSS is involved in purine synthesis while ARGSS is involved in arginine synthesis. Simulation showed that ADSS is consistently preferred over ARGSS, which is expected given the importance of purine synthesis in proliferation. On the other hand, many types of cancer are known to lack ARGSS and require external arginine [211]. Therefore, by applying global functional constraints on flux splits at metabolic branch points, biologically meaningful patterns can be identified.

This analysis is complementary to the approach in Chapter 4.2 and 4.3: while 4.2 and 4.3 focus on reaction pairs where one reaction is preferred over the other due to relative expression change between themselves, 4.4 focus on reaction pairs where one reaction is preferred over the other due to global network expression constraints.

4.5 Conclusion

Metabolic changes accompany many physiological (fasting and feeding) and pathological (diabetes and cancer) processes. Besides allosteric regulation, many metabolic changes are reflected at the transcriptomic level. Expression changes in metabolic affect the fate of key metabolites: over-expression of lactate dehydrogenase enables more lactate secretion from pyruvate, while over-expression of phosphoglycerate dehydrogenase diverts 3-phosphoglycerate from glycolysis to serine and glycine synthesis. This differential expression of metabolic genes results in differential allocation of key metabolites to downstream metabolic processes, which may have phenotype level consequences. At metabolic branch points, consistently higher expression of reaction i compared to reaction j (preferential state) across different samples of one phenotype might imply that reaction i is important to the metabolic state of the given phenotype. If this trend is consistently reversed between phenotypes (different preferential states), then the reaction pair and the associated metabolite may reflect key metabolic differences between the two phenotypes. By incorporating network-level expression constraints, additional reaction pairs with different preferential states between phenotypes can be identified, even though these reaction pairs may not show expression changes themselves. This approach is an important step to use mechanistic knowledge to increase our ability to identify perturbed genes and pathways.

4.6 Chapter 4 figures and tables

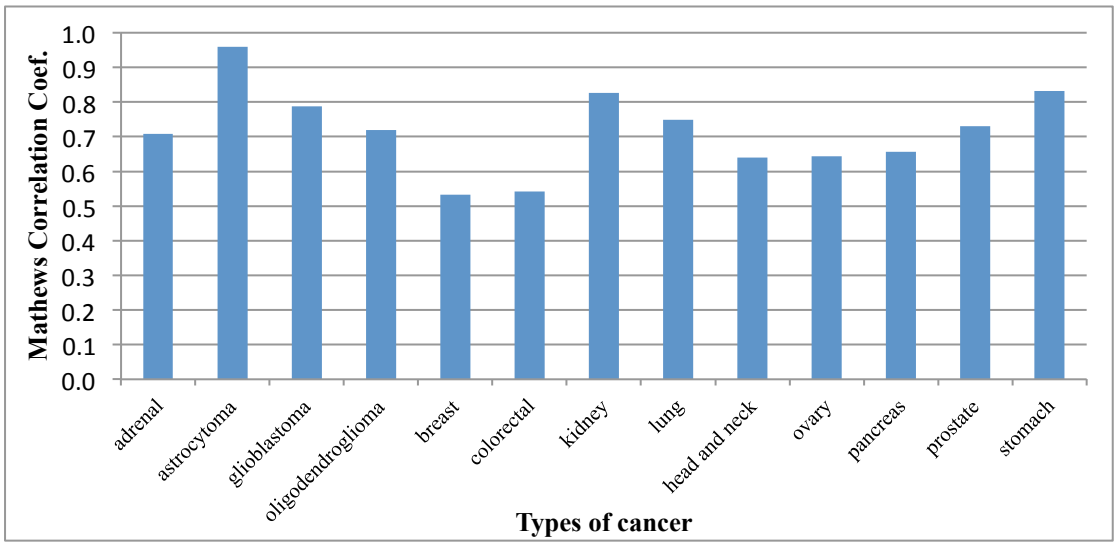

Figure 4.1 Predictive performances of metabolic branch point reaction pairs.

Reaction 1	Reaction 2	Metabolite	Adrenal	Astrocytoma	Glioblastoma	Oligodendroglioma	Breast	Colon	Kidney	Lung	Head and neck	Overy	Pancreas	Prostate	Stomach
S4T6g	S6T25g	Phosphoadenosine phosphosulfate	0.61	0.68	0.57	0.67	0.02	-0.02	0.96	-0.68	0.23	-0.15	0.25	0.03	0.22
ALCD2if	ALCD2yf	Ethanol	-0.61	-0.07	-0.09	0.00	-0.48	-0.60	0.39	-0.63	0.58	-0.55	0.19	-0.15	-0.62
GRTTx	IPDDIx	Isopentenyl diphosphate	0.30	0.46	0.84	0.32	-0.05	0.53	0.57	0.48	0.20	0.36	-0.03	0.18	0.57
S2T3g	S3T2g	Phosphoadenosine phosphosulfate	-0.18	0.75	0.78	0.71	-0.13	-0.34	-0.66	-0.22	-0.08	-0.44	-0.23	0.00	-0.35
GLCAT3g	S2T1g	Chondroitin sulfate C precursor 2	0.55	-0.66	-0.22	-0.56	0.39	0.38	0.31	0.53	-0.11	0.27	0.48	0.00	-0.05
ALCD22_D	PPDOy	D-lactaldehyde	0.00	0.12	0.02	0.06	-0.67	-0.66	0.21	-0.65	0.57	-0.57	0.19	-0.13	-0.65
S3T1g	S3T3g	Phosphoadenosine phosphosulfate	0.63	0.46	0.41	0.28	-0.26	0.47	0.44	0.23	-0.52	-0.12	0.43	-0.03	-0.18
CDS	CPCTDTX	CTP	-0.03	-0.68	-0.78	-0.71	0.16	-0.12	-0.38	0.17	-0.53	0.36	-0.25	0.10	-0.15
CATm	GTHPm	Hydrogen peroxide	-0.74	0.00	0.02	0.02	-0.49	-0.43	-0.49	-0.53	-0.07	-0.31	-0.43	-0.03	-0.60
ALCD22_L	LCADi	S-lactaldehyde	-0.10	-0.32	-0.22	-0.34	-0.61	-0.52	0.02	-0.40	0.64	-0.09	0.43	0.23	-0.21
ADSS	PRASCS	Aspartate	0.00	-0.88	-0.88	-0.82	0.08	-0.04	0.28	-0.37	-0.05	0.09	0.14	-0.20	-0.29
ARGSS	PRASCS	Aspartate	-0.45	-0.40	-0.41	-0.41	-0.39	-0.34	-0.55	-0.30	0.21	0.28	0.01	-0.15	0.18

Table 4.1. Metabolic branch point reaction pairs with consistent transcriptional reversals in multiple cancer v.s. normal tissue comparisions. Due to space limitation, reaction symbols are used. Corresponding reaction names and other information can be found in http://humanmetabolism.org/.

Reaction 1	Reaction 2	Metabolite	Adrenal	Astrocytoma	Glioblastoma	Oligodendroglioma	Breast	Colon	Kidney	Lung	Head and neck	Overy	Pancreas	Prostate	Stomach
GLGNS1	UDPGP	UDP-glucose	0.52	0.36	0.69	0.47	0.45	0.41	0.15	0.00	0.03	0.35	0.00	0.17	0.58
DOLPGT1_U	DOLPGT2_U	Dolichyl β-D-glucosyl phosphate	0.43	0.62	0.67	0.47	0.42	0.52	0.15	0.02	0.53	0.00	0.03	0.00	0.12
S2T3g	S6T3g	Phosphoadenosine phosphosulfate	0.55	0.68	0.66	0.69	0.45	0.00	-0.28	0.41	0.28	-0.19	0.38	0.00	0.26
BMTer_U	GPIMTer_U	Dolichyl D-mannosyl phosphate	0.43	0.32	0.18	0.38	0.41	0.54	0.33	0.36	0.32	0.63	-0.12	-0.35	0.21
GPIMTer_U	H3MTer_U	Dolichyl D-mannosyl phosphate	0.43	0.32	0.18	0.38	0.41	0.54	0.33	0.36	0.32	0.63	-0.12	-0.35	0.21
GTHS	PRAGSr	Glycine	0.43	0.27	0.64	0.50	0.48	0.33	0.15	0.21	0.30	-0.16	0.04	-0.06	0.44
GMPS2	NTD10	Xanthosine 5-phosphate	0.15	0.00	0.07	0.00	0.64	0.13	0.59	0.24	0.63	0.68	0.08	0.08	0.28
BPNT2	TYMSULT	Phosphoadenosine phosphosulfate	0.37	0.00	0.16	0.10	0.39	0.57	-0.36	0.29	0.27	0.67	0.20	0.00	0.69
BPNT2	ESTSULT	Phosphoadenosine phosphosulfate	0.37	0.00	0.16	0.10	0.39	0.57	-0.36	0.29	0.25	0.67	0.20	0.00	0.69
4NPHSULT	BPNT2	Phosphoadenosine phosphosulfate	0.37	0.00	0.12	0.10	0.39	0.57	-0.36	0.29	0.26	0.67	0.20	0.00	0.69

Table 4.2 Metabolic branch point reaction pairs with high entropy in cancer tissues but low entropy in normal tissues.

Reaction 1	Reaction 2	Metabolite	Adrenal	Astrocytoma	Glioblastoma	Oligodendroglioma	Breast	Colon	Kidney	Lung	Head and neck	Overy	Pancreas	Prostate	Stomach
CYSTA	PPNCL3	Cysteine	-0.16	-0.67	-0.55	-0.50	0.00	-0.28	-0.18	0.08	-0.31	0.00	0.00	0.00	-0.38
LCYSTAT	PCLYSOX	a-ketoglutarate	-0.67	0.00	0.00	0.00	0.00	-0.51	-0.65	0.00	-0.41	0.00	0.00	0.12	-0.24
PCLYSOX	TYRTA	a-ketoglutarate	-0.67	0.00	0.00	0.00	0.00	-0.51	-0.65	0.00	-0.41	0.00	0.00	0.12	-0.24
NNMT	SRTNMTX	S-Adenosyl-L-methionine	-0.22	0.00	0.00	0.17	0.00	0.00	-0.46	-0.34	-0.60	0.00	-0.51	0.00	0.00
CORE6GTg	N3Tg	Tn antigen	0.23	-0.41	-0.23	-0.21	0.00	-0.17	0.00	0.00	0.00	-0.23	-0.21	-0.18	-0.36
S3T3g	S4T6g	Phosphoadenosine phosphosulfate	0.14	-0.69	-0.40	-0.47	0.00	-0.23	-0.31	0.08	0.39	-0.19	-0.11	0.03	0.12
DOLPMT1_U	H8MTer_U	Dolichyl D-mannosyl phosphate	0.00	-0.48	-0.45	-0.42	0.00	-0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B3GALT44g	GALT2g	UDP-galactose	0.00	0.00	0.00	0.00	0.00	-0.64	0.00	0.06	-0.65	-0.34	0.00	0.00	0.00
ASNS1	PRASCS	Aspartate	0.23	-0.48	-0.58	-0.42	0.11	0.00	0.15	-0.05	0.06	0.09	-0.35	0.00	-0.33
ACITL	ACONT	Citrate	-0.46	0.27	0.00	0.10	-0.32	-0.17	-0.52	-0.04	-0.48	0.35	-0.42	-0.12	0.26

Table 4.3 Metabolic branch point reaction pairs with high entropy in normal tissues but low entropy in tumor tissues.

Reaction 1	Reaction 2	Pr(F_1>F_2 | Tumor)	Pr(F_1>F_2 | Normal)	Delta
ADPT	NNDPR	0.10	0.74	-0.63
GLUPRT	NNDPR	0.10	0.71	-0.61
ALCD2if	ETOHMO	0.92	0.39	0.53
GLYVESSEC	GNMT	0.64	0.21	0.43
AMETt2m	GNMT	0.92	0.50	0.42
GNMT	NORANMT	0.05	0.47	-0.42
ADSS	ARGSS	0.74	0.34	0.40
ARGSS	ASP1DC	0.26	0.66	-0.40
ARGtm	NOS1	0.82	0.42	0.40
ADMDC	GNMT	0.90	0.50	0.40
ARGSS	ASPCTr	0.26	0.63	-0.38
FPGSm	SARDHm	0.72	0.37	0.35
ALASm	GLYATm	0.44	0.11	0.33
CH25H	P45046A1r	0.28	0.61	-0.32
ARGN	NOS1	0.77	0.47	0.30
TRPHYDRO2	TRPO2	0.00	0.29	-0.29
P45046A1r	P4507A1r	0.79	0.53	0.27
AKGDm	SACCD3m	1.00	0.74	0.26
CH25H	P4507A1r	0.82	0.58	0.24
DHEASULT	PRGNLONESULT	0.23	0.03	0.20

Table 4.4. Metabolic branch point reaction pairs with large absolute Δ_r in pancreatic cancer.

Chapter 5. Reconstruction of genome-scale metabolic models for 126 human tissues

Human tissues perform diverse metabolic functions. Mapping out these tissue-specific functions in genome-scale models will advance our understanding of the metabolic basis of various physiological and pathological processes. The global knowledgebase of metabolic functions categorized for the human genome (Human Recon1) coupled with abundant high-throughput data now makes possible the reconstruction of tissue-specific metabolic models. However, the number of available tissue-specific models remains incomplete compared with the large diversity of human tissues.

In Chapter, 4, I report a method called metabolic Context-specificity Assessed by Deterministic Reaction Evaluation (mCADRE). mCADRE is able to infer a tissue-specific network based on gene expression data and metabolic network topology, along with evaluation of functional capabilities during model building. mCADRE produces models with similar or better functionality and achieves dramatic computational speed up over existing methods. Using our method, we reconstructed draft genome-scale metabolic models for 126 human tissue and cell types. Among these, there are models for 26 tumor tissues along with their normal counterparts, and 30 different brain tissues. We performed pathway-level analyses of this large collection of tissue-specific models and identified the eicosanoid metabolic pathway, especially reactions catalyzing the production of leukotrienes from arachidnoic acid, as potential drug targets that selectively affect tumor tissues.

This large collection of 126 genome-scale draft metabolic models provides a useful resource for studying the metabolic basis for a variety of human diseases across many tissues. The functionality of the resulting models and the fast computational speed of the mCADRE algorithm make it a useful tool to build and update tissue-specific metabolic models.

5.1 Background

Metabolic dysfunction has been implicated in a wide variety of human diseases such as obesity, diabetes, inborn errors of metabolism, neurodegenerative diseases, and cancer. The recent reconstruction of genome-scale models of human metabolism [123, 124] provides an important biochemical basis for systems analysis of metabolic related aspects of human physiology and pathology [125]. Such systems approaches are critical, as metabolism itself is a molecular transformation process where numerous metabolic pathways are inextricably interlinked[126]. However, the human body consists of many distinct tissues and cell types, each only expressing a fraction of the metabolic genes encoded within the genome [127]. Additional variability arises from environmental conditions and external stimuli. None of this variation can be fully accounted for with only the generic human metabolic model. Considering the context—e.g., genomic, anatomical, environmental, or temporal—under which a subset of the genome-scale biochemical network operates is therefore essential to understanding the molecular basis for many human diseases.

The importance of tissue-specific context in disease is evident from distinct metabolic profiles of cancers arising from different tissues. For example, it has been experimentally demonstrated that *MYC* oncogene-induced liver tumors show increased glutamine *uptake*, while *MYC*-induced lung tumors show glutamine *secretion* [29]. Another study showed that while lactate dehydrogenase A is important for breast carcinoma, neuroblastoma, and B-cell tumor cells, it is dispensable for *MYC*-induced lymphomagenesis [128]. Similar results were observed for phosphoglycerate dehyrogenease in breast cancer and melanoma [129, 130] versus *MYC*-induced lymphomagenesis [128]. Importantly, cancer metabolism in general also operates in unique environmental and signaling contexts compared to normal physiology and metabolic diseases such as obesity and diabetes[126].

The context in which a metabolic network operates can be viewed at multiple scales, all of which can be dependent on one another. The broadest level typically associated with metabolic models is genomic context—i.e., the full enzymatic capability encoded in the genome. Since the genome is the starting point from which to construct any generic organismal model, we will not consider it further

here. A more critical contextual consideration for genome-scale models in higher organisms—especially in human tissues—is the subset of metabolic enzymes that are being expressed (e.g., represented in the transcriptome) at a given time. The transcriptional regulatory state governs which subset of metabolic enzymes and pathways are active, and manifests as either (i) the specific expression program for a tissue or cell type; or (ii) the tissue or cellular response to intracellular or environmental conditions. The ideal strategy for modeling such contextual differences would be the integration of a generic, genome-scale model (e.g., *Human Recon 1* [123]) with a detailed, context-specific transcriptional regulatory network (TRN), including signaling events that relay cues from the cellular microenvironment. However, as these TRNs cannot yet be comprehensively and accurately reconstructed and modeled in human cells, recent efforts have turned to employing context-specific expression data to create models that are representative of active metabolism in specific human tissues and cell types either across a wide range of experimental conditions or under a particular condition [31, 131-139].

For clarity, we will henceforth delineate "*tissue*-specific" as meaning the representative active metabolic network for a tissue (e.g., liver, brain), and "*condition*-specific" as specific network states (e.g., hypoxia, drug treatment) of *tissue*-specific models. We also note that when higher resolution data is available, tissue-specific models can be further discretized into region or cell type specific models (e.g., different regions of brain, different neuron subtypes). Tissue-specific models are generally more desirable than condition-specific for predictive modeling, because they retain the flexibility and redundancy inherent in the metabolic network; specific conditions can subsequently be simulated directly by defining model constraints. Generating condition-specific models can still be highly useful, especially when coupled with experimental data for testing and validation; methods such as GIMME [137] and iMAT [131] have been used successfully to estimate the metabolic state of tissues under particular pathophysiological conditions. While the need for tissue-specific metabolic models is strong, the available number remains small. Importantly, significant knowledge and data is required to reduce a generic model to a tissue-specific model with enough rigors to allow for different condition-specific capabilities. Computational tools that can more rapidly generate tissue models that can represent a spectrum of physiological conditions will be highly useful for investigating metabolic dysfunctions in various diseases.

The Model Building Algorithm (MBA), a current state-of-the-art computational method to build tissue-specific metabolic models, has been used to build liver, generic cancer, and two cancer cell line metabolic models thus far [31, 135, 136]. The resulting models have been used to predict potential drug targets and improve metabolic flux predictions [31, 135, 136]. While a core set of high-confidence reactions in MBA is determined based on gene expression or literature evidence, the ranking and inclusion of non-core reactions is based on iterative model simulation for many different random reaction orderings. Notably, the random sampling in MBA—on the order of 1000 iterations in published studies—is limited in its coverage of the large space of possible orderings, potentially affecting the accuracy of the tissue-specific model. While this problem is mostly avoided by a stringent requirement in MBA for model consistency (i.e., all reactions in the final tissue-specific model must be capable of carrying flux), a more deterministic and simulation-independent ranking of non-core reactions would serve to dramatically speed up model construction time.

We have developed a method called metabolic Context-specificity Assessed by Deterministic Reaction Evaluation (mCADRE) that leverages gene expression evidence, network structure, and metabolic function to construct context-specific models in an automated, deterministic, and high-throughput fashion. Similar to MBA, mCADRE emphasizes the inclusion of a high-confidence core set of reactions from a generic genome-scale model, based on tissue-specific expression evidence. Non-core reactions are explicitly ranked according to their own expression evidence as well as weighted connectivity to other reactions in the network, and then sequentially removed in the inverse order of this ranking. The decision whether to confirm or reject each removal is determined by the consequent flux capacity of core reactions, as well as a universal test of metabolic functionality. To evaluate the performance of our method, we reconstructed a new liver model and compared results to liver models built by MBA: mCADRE was able to achieve similar coverage of high evidence reactions, improved metabolic functionality, and dramatic speed up. The deterministic decision making in mCADRE, coupled with an automated pipeline of data collection and processing, enables researchers to efficiently generate accurate and robust initial models from publicly available expression data.

As a demonstration of mCADRE's capabilities, we leveraged data from the Human Gene Expression Barcode Project[140] to automatically reconstruct draft genome-scale metabolic models for 126 human tissues and cell lines, collectively called the Tissue-Specific Encyclopedia of Metabolism (TSEM). All 126 metabolic models, mCADRE codes and input data are available at http://price.systemsbiology.net/downloads.php. We identified many amino acid metabolic pathways as enriched in 30 brain tissue models in TSEM, which agrees with the known role of amino acids in neurotransmitter metabolism. By comparing tumor and normal metabolic networks in TSEM, we also identified pathways with known roles in tumor metabolism. In particular, we identified part of the eicosanoid metabolic pathway as a potential selective target against tumor tissues. Further analysis of metabolic networks in TSEM, especially through integration with regulatory networks and various omics data, may offer novel insights of the metabolic aspects of various diseases.

5.2 Method overview and advantageous features of mCADRE

mCADRE builds a tissue-specific model from a generic human metabolic model [123] based primarily on gene expression data and metabolic network topology (**Figure 5.1**). Like MBA, we define a core set of reactions that should be present and active (i.e., able to carry flux) in the tissue model (we have implemented an adapted version of the *checkModelConsistency* module described in Jerby et al. to identify blocked reactions). The set of core reactions are determined from gene expression, and non-core reactions are evaluated and ranked according to a combination of expression and connectivity evidence (detail description in Material and Methods). To help ensure the basic functionality of the tissue-specific models, mCADRE includes a metabolic function test in the model building process. Specifically, the *checkModelFunction* module tests the ability of the current model to produce key metabolites from glucose, based on criteria previously used to universally evaluate such models [138]. This list can be customized based on literature evidence or metabolomics data (when available) to include tissue-specific metabolites or known capabilities of the tissue or cell type. We sequentially prune non-core reactions from the generic model in the determined order, provided that removal does not affect fluxes through the core reaction set or production of key metabolites from glucose. The former requirement is waived when removing non-core reactions whose associated

genes are not expressed in *any* tissue samples. For each reaction removed from the generic model, all resulting inactivated reactions are also removed.

5.2.1 Allowing for a flexible core reaction set increases tissue-specificity of metabolic pathways.

In addition to the core set of reactions (whose associated genes are expressed in many tissue samples), mCADRE defines a negative set of reactions whose genes are not expressed in any tissue samples. In this case, when expression evidence strongly suggests that a reaction should not be included, we relax the constraint described above to also allow for removal of any consequently inactivated core reactions. The non-expressed reaction is removed, along with all reactions that can no longer carry flux, only if the ratio of resulting inactivated core reactions to inactivated non-core reactions is smaller than a specified ratio. This parameter governs the sensitivity versus specificity of the final tissue model: a lower ratio cutoff leads to inclusion of more reactions with strong positive evidence, while a higher cutoff leads to removal of more reactions with strong negative evidence. It is important to note the difference between non-expressed reactions (expression evidence strongly suggest the absence of such reactions) and non-gene associated reactions (no expression evidence available). Non-gene associated reactions include spontaneous reactions and reactions catalyzed by enzymes not annotated to genes yet. No core reactions are allowed to be removed when mCADRE tries to remove non-gene associated reactions.

The utility of allowing for a flexible core can be seen with the bile acid biosynthesis pathway in the liver. Although many cell types express several enzymes in the bile acid pathway, the complete pathway is present only in the liver [141]. The tissue-specificity of this pathway is also supported by microarray data: among 126 tissues, we found that almost all bile acid synthesis reactions have strong evidence of activity in the liver, but not in other tissues (**Figure 5.2**). However, a few reactions in this pathway have strong evidence in non-liver tissues (e.g., cerebral cortex). If a "hard" core were to be enforced—i.e., requiring all reactions with expression evidence above a threshold to carry flux and remain in the tissue model—most reactions in the bile acid synthesis pathway would be included in these tissues, even though most reactions in the pathway lack expression evidence. When we allowed for a flexible core, only liver and liver cancer models included almost complete bile acid synthesis

pathways (85% of pathway reactions present), while most other tissues did not have reactions from this pathway. In contrast, when using a hard core, most models included a majority of bile acid reactions (60%~80% of pathway reactions), conflicting with known tissue specificity. For cerebral cortex, for example, 70% or 5% of the bile acid synthesis pathway is computed as present when using a hard or flexible core, respectively, supporting the importance of including the flexible core in the mCADRE approach. This result is achieved when the inactivated core to non-core reaction ratio is set at 0.33, and is robust to the ratio (Table 5.1).

There are 541 and 844 confidently positive (expressed in more than 50% of tissue samples) and negative (not expressed in any tissue sample) metabolic reactions in the cerebral cortex, respectively. 49 confidently negative reactions are needed for the basic functionality of the model (glycolysis, TCA cycle, pentose phosphate pathway). To remove all the other 795 confidently negative reactions, 185 (34.2%) confidently positive reactions have to be removed. On the other hand, to include all 541 confidently positive reactions, 231 (27.37%) confidently negative reactions have to be included. At ratio 0.33, as shown in the table 4.2, 5.18% confidently positive reactions are removed, while 79.62% confidently negative reactions are removed.

There are 172 reactions whose associated metabolic genes are not expressed in any cerebral cortex microarray samples but still retained in the model. 130 of the 172 reactions' associated metabolic genes have protein-level staining evidence available in the Human Protein Atlas(HPA) [142]. Among these 130 reactions, according to gene-reaction mapping, 21, 69, 36, and 4 reactions have negative, weak, moderate and strong protein staining evidence, respectively. Therefore, more than 80% of confidently negative reactions based on transcriptomic data do have non-negative proteome-level evidence. This can either be caused by limitations of the microarray geneexpression measurement, or potential effects of post-transcriptional regulation of the corresponding metabolic genes.

On the other hand, compared with enforcing a hard core reaction set, at ratio 0.33, 28 core reactions are removed. Among the 28 reactions, there are 5 UDP-glucuronosyltransferase reactions encoded by UTG1A8 and UTG1A10 (gene-to-reaction rule: UTG1A8 or UTG1A10). UDP-glucuronosyltransferase reactions occur mainly in the liver and intestines, and neither UTG1A8 nor UTG1A10 are expressed in brain in general or cerebral cortex in particular[143, 144]. 3 reactions

in bile acid synthesis, all associated with the SCP2 (sterol carrier protein 2) are also removed. Sterol carrier protein 2 shows negative staining in cerebral cortex neuronal and glial cells in HPA. However, according to the input microarray data from the Gene Expression Barcode, all UDP-glucuronosyltransferase reactions and reactions associated with sterol carrier protein 2 have high expression-based evidence (0.65 and 1, respectively).In these 2 examples, mCADRE correctly removed these reactions because many reactions with little expression evidence are needed to maintain flux through these reactions, and mCADRE tries to balance strongly positive and negative expression evidence.

5.2.2 mCADRE significantly reduces computation time to generate context-specific models

The novel reaction ranking scheme based on three criteria (gene expression, network connectivity, and literature-supportted reaction confidence level encoded in *Human Recon 1*) enables mCADRE to perform a single optimized iteration to infer a tissue-specific model from the generic human metabolic map. In contrast, MBA determines whether to retain non-core reactions through a large number of random iterations (typically ~1000) to account for the effects of the order in which reactions are removed [135]. The order of reaction removal remains influential in mCADRE—e.g., redundant reactions can be removed interchangeably with equal effects on core reactions, but whichever reaction is pruned first mandates the retention of the latter. However, mCADRE leverages gene expression, topology, and literature evidence to directly determine ordering in a quick deterministic fashion, avoiding random iterations. Each iteration in MBA involves Flux Variability Analysis (FVA [145]; maximization and minimization of all reactions to calculate flux capacity), which amounts to on the order of 10,000 separate optimizations. This computational complexity limits not only the throughput of reconstructing metabolic models, but potentially their reproducibility: as the full solution space of ordered reaction removals is extremely large, the 1000 permutations sampled by MBA necessarily represent only a tiny fraction of all possibilities. The stringent requirement in MBA that final tissue-specific models be consistent (i.e., contain no gaps) enforces the inclusion of many lower-evidence reactions, and leads to mostly similar models from run to run. Still, even with a heuristic speed-up [135] or the efficient fastFVA algorithm [146], one iteration of MBA takes ~10 hours on a single 2.34 GHz CPU with 4G RAM using the open source glpk solver. The whole MBA reconstruction process, with ~1000 iterations, would therefore take on the order of ~10,000

CPU-hours. mCADRE dramatically improves the computational speed via deterministic evidence-based evaluation of reactions, requiring only ~10 CPU-hours under the same configuration. With the IBM Cplex solver (free for academic institutions), the model reconstruction time is further reduced to 4 hours.

Manual curation of metabolic models often continues over several iterations of simulation-based hypothesis generation, experimental validation, and model refinement to improve quality and predictive accuracy. Such iterative curation is also important in computational model reconstruction, especially when new and better (i.e., more comprehensive, more sensitive, higher resolution) data becomes available. New technologies such as RNA-seq provide unprecedented characterization of the transcriptome [147] with a much lower detection limit than microarray. As RNA-seq data become available for a variety of tissues and cell types [148], it is important that corresponding models are updated to better reflect the metabolic capacity corresponding tissues: metabolic genes expressed at low levels may be regarded as not expressed by microarray and excluded from metabolic models. Because mCADRE reduces the computational time of model reconstruction almost 1000 fold, it is much more convenient to build or update a large collection of tissue-specific models when new data are released.

5.3 Coverage-based and functional validation of a mCADRE-constructed liver model

As initial validation of the mCADRE method, we used the algorithm to reconstruct a liver model (*liver*CADRE) and compared it to the liver model in the original MBA publication [135] (henceforth referred to as *liver*MBA), as it is the best characterized MBA-generated tissue model to date. We built the *liver*CADRE model based on 23 normal liver microarray samples. Notably, both mCADRE and MBA result in consistent final tissue models, so all liver model reactions examined are able to carry flux (**Table 5.1**). While *liver*CADRE includes 1763 reactions to the 1826 in *liver*MBA, the two models share 1473 reactions, a significant overlap (Under hypergeometric distribution, the probability of observing 1473 or more overlapping reactions is 1.54×10^{-12}, N = 2469, the total number of flux-carrying reactions in *Recon 1*); these overlapping reactions constitute over 80% of all reactions in each model, and thus substantial convergence between the approaches, establishing confidence for the

quality of models generated with mCADRE. To more directly evaluate the performance of mCADRE and MBA for generating new tissue-specific models, we also used MBA to build a model from our liver expression training data; *liver*CADRE exhibited similar or better coverage and increased functionality in comparisons with the new MBA model built with the same training data.

5.3.1 mCADRE-constructed liver model improves coverage of highly expressed genes and proteins

While the two models share most reactions, we chose to further explore the gene-associated reactions unique to each model. There are 194 and 169 gene-associated reactions unique to *liver*CADRE and *liver*MBA, respectively. For each set of reactions, we first examined the coverage of highly expressed metabolic genes in an independent data set (test data set), not used in building either liver model and based on a different microarray platform than any of the training data used by mCADRE. We assume that reactions with strong expression-based *validation* (whose associated metabolic genes are most ubiquitously expressed across new tissue-specific samples) are more likely to be present in the liver. The set of gene-associated reactions unique to the mCADRE model have higher expression-based validation score than gene-associated reactions unique to *liver*MBA (Wilcoxon rank sum test p-value: 6.02×10^{-9}; **Figure 5.3A**). *liver*CADRE includes more reactions with strong expression-based validation than the MBA model (46% vs. 18%) and fewer reactions with low scores (i.e., poor to no gene expression-based validation) than *liver*MBA (47% vs. 76%; **Figure 5.3A**).

As mRNA and protein levels are only moderately correlated in mammalian cells [149], we also compared coverage of the two liver models at the protein expression level. We collected protein staining data from the Human Protein Atlas (HPA) [150] for 560 metabolic genes in *Recon 1*. Protein staining strength is divided into four levels by HPA: strong, moderate, weak, and negative. We mapped the HPA data to reactions according to gene-reaction rules, and assigned each reaction a score, based on the staining strength of its associated metabolic gene. As shown in **Figure 5.3B**, 29% of gene-associated reactions unique to *liver*CADRE have strong protein staining support, and 20% have negative staining support. In comparison, only 4% of reactions unique to *liver*MBA have strong protein staining support, while 40% have negative protein staining support. Moreover, gene-associated reactions unique to *liver*CADRE collectively have significantly higher scores than *liver*MBA (Wilcoxon rank sum test *p*-value: 1.25×10^{-5}).

5.3.2 mCADRE-based liver model is functionally comparable to the existing liver model

The liver plays a major role in metabolism and carries out important metabolic functions such as gluconeogenesis, triglyceride synthesis, amino acid degradation, and ammonia and ethanol detoxification. We investigated the ability of the two liver models to carry out these hepatic metabolic functions; the details of each metabolic function simulation are described in Materials and Methods. We found that the two models are functionally comparable (Table 5.3), which is expected given that they share a large percentage of reactions. Both models can detoxify ammonia and ethanol; both can simulate gluconeogenesis from physiologically important substrates such as pyruvate, lactate, alanine and glutamine; and both models can degrade most amino acids and produce urea as byproduct. While *liver*MBA can degrade more amino acids and generate glucose from a broader range of glucogenic substrates, only *liver*CADRE is able to synthesize triglyceride from glucose and fatty acids. Triglyceride synthesis is a major hepatic function underlying blood glucose and lipid homeostasis—the ability to simulate this function *in silico* enables the investigation of liver metabolic network states in normal and pathological conditions such as obesity and fatty liver disease. While *liver*MBA included over 700 reactions manually curated to be active in the liver[135], no such curation is done to build *liver*CADRE. *liver*CADRE also outperforms the new MBA liver model (*liver*MBA^{wang}, to distinguish from the original MBA model) built with the same training data in liver metabolic function tests.

The liver is also able to regenerate after injury, which involves the synthesis of biomass precursors such as nucleotides, amino acids, and lipids. A biomass reaction was added to both models (after construction) that accounts for the growth requirement of amino acids, nucleotides, lipids, and other metabolites and we tested the ability of the two models to grow *in silico* in RPMI 1640 tissue culture medium conditions. The *liver*CADRE model was able to simulate growth without further manual curation, while *liver*MBA lacked this capability. Further analysis identified that *liver*MBA could not grow because it contained no reactions in the inosine monophosphate (IMP) pathway, and therefore could not produce purines. As *de novo* purine synthesis primarily occurs in the liver [141], this lack of this capability represents a metabolic gap in *liver*MBA. Moreover, many membrane phospholipids such as phosphatidic acid, phosphatidylethanolamine, phosphatidylcholine and phosphatidylserine are derived from triglyceride. As *liver*MBA cannot produce this metabolite, the production of these

glycerophospholipids is consequently blocked. This demonstrated the importance of the metabolic function test in mCADRE, as it ensures the basic functionality of the resulting model and may save substantial *post hoc* manual curation.

5.4 mCADRE for high-throughput model generation

After verifying that our new liver model could show similar or better coverage and functionality when compared to state-of-the-art models and algorithms, we next took advantage of the automated and computationally efficient nature of mCADRE to generate a large collection of tissue and cell type specific metabolic models. The Gene Expression Barcode project previously collected, annotated and binarized microarray data for 126 human tissues and cell lines on the Affymetrix U133Plus2 platform [140]. We used these binarized microarray data sets as input evidence in mCADRE to extract individual models from the generic *Recon 1*, thereby establishing a Tissue-Specific Encyclopedia of Metabolism (TSEM). This effort provides the most comprehensive mapping to date of human tissue-specific metabolic networks, and for many of the 126 tissues or cell types, this represents the first time a genome-scale metabolic model has been built.

5.4.1 Tissue-specific Encyclopedia of Metabolism enables global analysis of human tissues

The Tissue-Specific Encyclopedia of Metabolism (TSEM) includes 26 tumor tissues and cell lines, and 18 of these tumor tissues also have corresponding normal tissue models. It also contains metabolic models of 30 different brain tissues, many of which are affected in various neurological diseases. A full list of the 126 tissues and the corresponding microarray data can be found at [140] and http://rafalab.jhsph.edu/barcode/. All new metabolic models already include several important features for *in silico* simulation of cellular behavior: they have functional central metabolic pathways (glycolysis, TCA cycle, pentose phosphate pathway), can synthesize non-essential amino acids and nucleotides from glucose, have functional fatty acid synthesis pathway (from acetyl-CoA to palmital-CoA), and can synthesize key membrane lipids. These functionalities are a result of the basic universal metabolic function test integrated into mCADRE, which can be further customized to reflect the specific capabilities of individual tissues and cell types. The latest version of these models, as well

as additional models built with latest data (e.g., RNA-seq) can be downloaded from http://price.systemsbiology.net/downloads.php

With this comprehensive set of tissue-specific draft metabolic models, we can start to evaluate global properties of these networks and their relationship to human metabolism in the body (Figure 5.4). Models in the TSEM contain 1161 reactions on average (47% of flux-carrying reactions in *Recon 1*), with most models ranging from 1000 to 1300 reactions (Figure 5.4A). The smallest model is neutrophiles, which included 826 reactions. The largest models are liver tumor, normal liver and kidney, which included 1550, 1530 and 1416 reactions respectively. This is expected as the liver and kidney are among the most metabolically active tissues in the human body. There are 2311 reactions that appeared in at least one of the 126 context-specific models, representing 93% of the flux-carrying reactions in *Recon 1* (Figure 5.4B); 600 reactions appear in at least 90% of the 126 models, and 546 reactions appear in at most 10% of the models.

5.4.2 Distributions of TSEM model reactions correspond to known features of brain and tumor tissues

We identified pathways that are enriched in brain tissue models—i.e., pathways with more reactions present in normal brain tissues than in normal non-brain tissues (Table 5.4). Among the pathways most enriched for completeness in brain (compared to normal non-brain tissues) were taurine and hypotaurine metabolism, aromatic amino acid biosynthesis, cysteine metabolism, alanine and aspartate metabolism , glutamate metabolism, and valine, leucine, and isoleucine metabolism. This makes sense, as many amino acids are either neurotransmitters or intermediates in neurotransmitter synthesis. The brain is also known to contain higher concentrations of long-chain polyunsaturated fatty acids (PUFAs) than most other tissues, and both cerebral endothelium and astrocytes elongate and desaturate precursors of the long-chain PUFAs [151, 152]. As such, it is not surprising to see that the fatty acid elongation pathway includes significantly more reactions in brain tissues than non-brain models. Pathways better enriched in brain tissue models are in agreement with known brain-specific metabolic functions, demonstrating the quality of these models.

We also identified pathways enriched in the 18 tumor tissues compared to their 17 corresponding normal tissues (including two different tumors that arise from the same normal tissue; Table 5.5), including folate metabolism, eicosanoid metabolism, fatty acid activation and nucleotide metabolism. Folate metabolism is necessary for *de novo* nucleotide synthesis. The enrichment of reactions for this pathway—as well as the nucleotides pathway—in tumor tissue models makes sense because nucleotide synthesis is more active in proliferating tumor cells, and many enzymes in nucleotide synthesis are classical chemotherapy targets.

Additionally, tumors overexpress fatty-acid synthase (*FASN*) and undergo significant *de novo* fatty-acid synthesis [153]—*FASN* has been identified as a drug target in many tumors [154]. Fatty acid activation reactions are catalyzed by acyl-CoA synthetase (*ACS*), which acts downstream of FASN and converts long-chain fatty acids to acyl-CoA. Fatty acid activation is a critical step in several lipid metabolic pathways, including phospholipid and triacylglycerol biosynthesis. Some genes in this pathway (e.g., *ACSL4* and *ACSL5*) are overexpressed in certain types of cancer and inhibition of these genes induced apoptosis in cancer cells [155]. Notably, eicosanoid metabolism is the second most tumor-enriched pathway. Eicosanoids, which are biologically active lipids derived from arachdonic acid by cyclooxygenase, lipoxygenase, and P450 epoxygenase, have been implicated in inflammation and cancer [156]. Biologically active sphingolipids are involved in cancer pathogenesis-ceramide functions as a tumor-suppressor lipid, while sphingosine-1-phosphate functions as a tumor-promoting lipid[157]. This supports the identification of the sphingolipid pathway as enriched in tumor metabolic networks.

As a comparison, we also calculated the enrichment statistic of these brain and tumor enriched pathways using only expression data. Using the gene-reaction-pathway annotation in Recon 1, we calculated the average ubiquity score (i.e., how often a metabolic gene is expressed in tissue samples) of metabolic genes in a pathway, and compared the average ubiquity scores of the above pathways in brain vs. non-brain and tumor vs. normal tissues. As shown in Table S12, only 4 of the 10 brain-enriched pathways and 1 of the 9 tumor-enriched pathways identified by the model-based approach are also found by expression data alone, respectively. This shows the increased signal we can get through the model-based approach.

We repeated the analysis to identify *individual reactions* that occur significantly more frequently in tumor tissue models than in normal tissues. Interestingly, the top most differentially included reactions (Table 5.6) together form part of the eicosanoid metabolism pathway, from arachidonic acid to leukotriene A4, C4, D4, E4 and F4 (Figure 5.5). The first two reactions are catalyzed by 5-lipoxygenase, which is induced by inflammatory stimuli and is often constitutively expressed in various cancers [156]. Furthermore, inhibition of 5-lipoxygenase has been shown to reduce cell proliferation and angiogenesis [158] and augment the antitumor activity of other drugs [159]. Leukotrienes have been implicated in various diseases such as asthma, cardiovascular diseases, and cancer [160]. For example, leukotriene C4 and D4 promote angiogenesis [161]; leukotriene D4 also promotes intestinal epithelial cell migration [162]. While involvement of genes and metabolites in the eicosanoid metabolic pathway has been reported in some cancers, our pathway and reaction level analysis revealed the importance of this pathway across a broad range of tumors arising from many different tissues.

5.4.3 Comparision of TSEM kidney model with the existing kidney metabolic model

As a demonstration of the utility of models in TSEM, we compared the TSEM kidney (*kidney*TSEM) metabolic model with the existing reduced kidney metabolic model (*kidney*Reduced) from Chang *et al*[133]. While *kidney*TS EM is a genome scale metabolic model with 1416 reactions, *kidney*Reduced aims to capture ony the core kidney metabolic phenotype and only included 443 reactions. First, we compared the 578 and 34 gene-associated reactions unique to each model that also have protein staining evidence (Table 5.7). 554 and 32 of gene-associated reactions unique to each model have non-negative protein staining, respectively. Thus, the *kidney*TSEM model included much more reactions that are active in the kidney.

Chang *et al* compiled a list of 41 important renal metabolic functions, which consists of the secretion and adsorption of metabolites involved in blood pressure. Following Chang *et al*, we added exchange and demand reactions for the 41 metabolites, and maximized the adsorption or secretion of each metabolite according to whether it is adsorbed or secreted by the kidney.

*kidney*TSEM is able to achieve 35 of the 41 renal metabolic functions, while *kidney*Reduced achieved all 41 renal metabolic functions. The 6 renal functions that kidneyTSEM failed to achieve includes the

secretion of prostaglandin I2, vitamin D, tryptamine, and the absorption of acetate, oxalate, and L-carnosine.

Chang *et al* also compiled a list of 20 gene deficiencies that are known to cause kidney disorders. 11 of the 20 genes are in *kidney*TSEM model, and all 11 gene deficiencies are predicted by *kidney*TESM to affect at least one of the 35 renal metabolic functions *kidney*TSEM can achieve under normal conditions. Thus, the recall rate is 55%, but precision is 100%.

Although not as good as *kidney*Reduced, *kidney*TSEM can simulate most renal metabolic functions and have good predictability of genetic perturbations. It is important to note that while *reduced*Kidney is based on significant amount of metabolomics data, transcriptomic data, and most importantly, manual literature curation, the reconstruction of *kidney*TSEM is fully automated using transcriptomic data.

5.5 Comparing mCADRE with the recently published INIT method

During the preparation of this manuscript, a new method (Integrative Network Inference for Tissues, INIT) capable of genome scale tissue-specific metabolic network reconstruction was published. This method was used to build genome-scale metabolic networks for 69 human cell types and 16 cancer types, collectively referred to as the Human Metabolic Atlas(HMA) [25]. The fundamental strategy of mCADRE and INIT is somewhat similar. Both mCADRE and INIT start from a generic human metabolic model, and use expression data to infer a tissue-specific sub-network. Both methods require that the model should be able to produce certain important metabolites. While mCADRE requires the model to produce universally important metabolites from simple precursors like glucose (which may overestimate the metabolic capabilities of human cells), INIT allows the model to uptake all metabolites to produce these key metabolites (which may underestimate the metabolic capabilities of human cells).

There are certain important differences between the two methods. INIT primarily uses the evidence from the Human Protein Atlas (HPA) as input, but can also incorporate gene expression and metabolomics data; algorithm parameters (e.g., weights assigned to different level of evidence) are also optimized for HPA data [25]. mCADRE can use any data that can quantitatively measure mRNA

or protein abundance, but herein primarily uses gene expression microarray data. Although proteome data from HPA provide more direct evidence for the existence of corresponding metabolic reactions, current proteomic data is much less comprehensive than transcriptomic data from microarray or RNA-seq. To date, transcriptomic data from public repositories have explored a much broader range of human tissue/cell types and pathophysiological conditions: we were able to use mCADRE to build 126 human tissue-specific metabolic models based on transcriptomic data from a single microarray platform. Additionally, while mCADRE and other model building algorithms (MBA, GIMME, iMAT, etc.) are based on the steady state assumption of no net accumulation of metabolites, i.e., mass-balance, INIT allows for a small positive net accumulation of metabolites if a metabolite is present in a cell type according to metabolomics data.

We validated mCADRE by showing that the mCADRE-built liver model can simulate a wide range of liver metabolic functions. An INIT-built liver model was shown to more accurately cover liver metabolic gene expression than the manually reconstructed HepatoNet1 [134]. However, it is unclear how this better coverage at the gene level will translate to model functionality: while HepatoNet1 was tested to be able to simulate a comprehensive set of 442 liver metabolic objectives, no such metabolic function simulation was reported for the INIT-built liver model.

We also compared the Tissue-Specific Encyclopedia of Metabolism (TSEM) built by mCADRE and the Human Metabolic Atlas built by INIT. Overall, TSEM included 126 models and HMA included 85 models. HMA included 16 cancer models while TSEM included 26 cancer models; 11 are shared. There are 100 and 69 normal tissue/cell types in TSEM and HMA respectively. Overall, the two collections shared 21 normal tissues (counting multiple cell types of the same tissue in HMA as one single tissue), with 30 unique to HMA and 79 to TSEM. One main difference is the coverage of tissues in the brain. While HMA included 8 models covering 3 brain structures (each brain structure has 2 or 3 cell-type specific models), TSEM included 30 models covering 30 distinct brain structures. Thus, the two collections of tissue-specific metabolic models are largely complimentary, with TSEM covering many more tissues.

5.6 Conclusion

Large amounts of data have accumulated in public data repositories, characterizing the molecular phenotype of a variety of human tissue and cell types across a wide range of pathophysiological states [163]. However, the number of available tissue-specific metabolic models, which enable the systematic simulation of metabolic functions in normal and disease contexts, remains relatively small. To bridge this gap, we have developed a new automated method (metabolic Context specificity Assessed by Deterministic Reaction Evaluation, mCADRE) to efficiently build tissue-specific metabolic models in a high-throughput manner. From the comparison of brain and non-brain tissue models and the comparison of tumor and normal tissue models, it is clear that the pathway-level analysis is in agreement with literature. The corresponding models therefore enable further exploration of brain-specific metabolic functions and identification of drug targets that specifically kill tumor cells with minimal side effects on normal tissues. Combined with automated data acquisition and annotation tools [164, 165], mCADRE has the potential to transform large repositories of gene expression data into repositories of functional tissue-specific metabolic models.

Importantly, metabolism is under extensive transcriptional regulation [166]. Mutations in transcription factors can cause various metabolic diseases [167], and many tumor suppressor genes and oncogenes are also transcriptional regulators of metabolism [27]. Combined with methods that automatically integrate transcriptional regulatory networks and metabolic networks [85], mCADRE may help to systematically identify the metabolic effects of transcription factors perturbations in various tissues. Ultimately, we hope to expand the TSEM to include both metabolic and corresponding transcriptional regulatory networks for many tissues and cell types. Additionally, metabolic interactions *between* different tissues and cell types play important roles in health and disease [168-171], and there have been pioneering studies that used integrated multi-cell type or multi-tissue type models to such interactions [138, 139]. The large collection of tissue and cell type specific models in the TSEM may facilitate the integrated modeling of metabolic interactions such as those between adipocytes and macrophages, different brain tissues, and between tumor and stromal microenvironment.

5.7 Materials and Methods

The majority of the automated reconstruction pipeline in mCADRE, including the MAS5 detection call, is implemented in Matlab, and the pipeline produces genome-scale draft metabolic models from raw expression intensity files. To validate this pipeline, we built a liver model with MAS5 as the binarization method and compared it to a liver model constructed with MBA. Aside from the hepatic functional testing of liver models, all steps described below were subsequently applied to generate context-specific models for 126 different human tissues. Note that the Gene Expression Barcode project already used the barcode method to produce binarized transcriptomic data for the 126 tissues, so MAS5 was not used in this case. As the barcode binarization tends to be more stringent in calling a gene expressed than MAS5 [140], the resulting models may also be smaller than when MAS5 is used.

5.7.1 Gene expression data processing

This new method uses gene expression microarray data as input evidence to prune a generic model (e.g., *Human Recon 1*) to a context-specific subset; the SBML file for *Recon 1* was obtained from the BiGG database [172] and converted into COBRA Toolbox [173] model structure for subsequent analysis. To construct a context-specific metabolic model for the liver, we acquired raw gene expression profiles from 23 liver tissue samples from [116, 174-176] and GSE7307 (no citation available). All of these studies were identified and annotated by the Gene Expression Barcode Project [140]. To approximate the presence or absence of the enzyme and transporter-encoding gene in a particular profile, we used the Affymetrix MAS5 detection call to binarize raw microarray data [177]: present calls are treated as 1, while marginal and absent calls are treated as 0. Other binarization methods, such as the gene expression barcode [140], can also be used. The final binarized expression data for all genes $g \in G$ in samples $n \in N$ for a selected context or phenotype is represented as the expression matrix $X^{|G|\times|N|}$, where $X_{g,n} = \{0,1\}$ represents the presence of gene g in sample n; for our liver training data, $|G| = 20{,}283$. There are 54,613 probe sets on the Affymetrix U133Plus2 platform (excluding quality control probes). Only probes that can uniquely map to a single gene are retained; these probes map to 20,283 unique genes. When multiple probes map to the same gene, the maximum expression value is used.

5.7.2 Assigning evidence scores to reactions

For each reaction $r \in R$ in the generic model, we assign evidence scores ($\boldsymbol{E}(r)$) to deterministically evaluate which reactions to keep or remove when pruning to get a context-specific network. We first calculate the *expression-based evidence* $\boldsymbol{E_x}(r)$ for all reactions to provide an overall ranking and to divide reactions into core and non-core sets. Next, the network topology of the generic model is used to calculate the *connectivity-based evidence* $\boldsymbol{E_c}(r)$ for each non-core reaction; this provides a second level of evidence when determining the order of reactions to remove during pruning. Finally, if expression- and connectivity-based evidence is insufficient to determine the rank of a reaction, *evidence based on confidence level* in the generic model $\boldsymbol{E_l}(r)$ is considered.

Expression-based evidence

After binarizing the input data, we first quantify how often a gene is expressed across samples of the same context; this is the ubiquity score $U(g)$ for each gene g:

$$U(g) = (1/|N|)\sum_{n \in N} X_{g,n}.$$

This score ranges from 0 (not expressed in any context samples) to 1 (ubiquitously expressed in context samples). According to gene-reaction rules, ubiquity scores for metabolic genes are mapped to corresponding reactions. That is, the expression-based evidence $\boldsymbol{E_x}(r)$ for reaction r is a function of how often its associated genes $g_r \in G_r$ are expressed in the selected context, as measured by the ubiquity score:

$$\boldsymbol{E_x}(r) = f(U(g_r)),\ g_r \in G_r.$$

The relationship between the ubiquity scores of G_r and $\boldsymbol{E_x}(r)$, denoted by f, is a composite of the Boolean gene-reaction rules defined in the generic model: AND is replaced with MIN, while OR is replaced with MAX, following ref [137] (**Figure 1A**). By definition, the expression-based evidence

$\boldsymbol{E_x}(r)$ also ranges from 0 to 1, indicating how likely the reaction is to be present in the selected context. The high-confidence core set of reactions is then defined as those with $\boldsymbol{E_x}(r) > 0.9$ when building *liver*CADRE with MAS5 call binarization, and $\boldsymbol{E_x}(r) > 0.5$ when building the 126 tissue models with the Barcode binarization. Higher cutoff is used for MAS5 call binarized data, as it is less stringent than Barcode in calling a gene expressed. Reactions with $\boldsymbol{E_x}(r) = 0$ are defined as the negative reaction set: these reactions have strong evidence of not being active in the tissue context.

Connectivity-based evidence

For non-core reactions, we use network topology to define a secondary metric called *connectivity-based evidence* $\boldsymbol{E_c}(r)$. This score is particularly designed to rank non-gene-associated reactions, which account for 40% of all reactions in *Recon 1*, and by definition, will not be in the core because they are not associated with expression data. The connectivity-based evidence for each non-core reaction accounts for both the expression-based evidence and connectedness of all adjacent reactions (core or non-core). Using the stoichiometric relationships defined in the *S* matrix, we can describe whether any two reactions in the generic model are connected (i.e., share at least one metabolite) with the binary *adjacency* matrix $A^{|R|\times|R|}$. Specifically, $A_{i,j} = \{0,1\}$, where 1 indicates reaction *i* is connected to reaction *j*.

We consider the outgoing *influence* $I(r)$ of each reaction as its normalized connectedness to all adjacent reactions. That is, for each reaction *r*,

$I(r) = 1/\sum_{j\in R/r} (A_{r,j})$.

In this way, *r* exhibits influence on all other reactions $j \in R/r$ that is inversely proportional to the number of reactions to which it is connected. Furthermore, we measure the *weighted influence* $WI(r) = \boldsymbol{E_x}(r)\times I(r)$ such that *r* will exhibit stronger influence on connected reactions if it was found to have strong expression-based evidence; reactions with $\boldsymbol{E_x}(r) = 0$ thus have no weighted influence on adjacent reactions.

Finally, we define connectivity-based evidence $\boldsymbol{E_c}(r)$ as the net incoming weighted influence to reaction r from all other reactions $j \in R/r$:

$$\boldsymbol{E_c}(r) = \sum_{j \in R/r} (WI(j) \mid A_{r,j} = 1).$$

If a non-core reaction r_j is connected to a highly expressed reaction r_i that has few other connections, this provides strong support for its inclusion in the context-specific model. Conversely, if a core reaction r_i is connected to many other reactions, then it is less clear whether any particular connected non-core reaction r_j is the one that functions in a pathway with r_i in the pruned network; as such, the resulting connectivity-based evidence for r_j will be lower.

Confidence level-based evidence

Confidence scores indicate the level of biological evidence associated with each reaction, as determined during manual curation of the generic metabolic model—in this case, *Human Recon 1*. The confidence level evidence $\boldsymbol{E_l}(r)$ for a reaction ranges from 1 (*in silico* modeling evidence only) to 3 (experimental biochemical or genetic evidence); midlevel scores (2) indicate some physiological evidence, or experimental support from a related organism, and a score of 0 indicates that the reaction was not evaluated. Importantly, these confidence scores represent evidence for the generic model, not for the specific context, and thus are considered as a tertiary measure of evidence for non-core reactions.

5.7.3 Pruning the generic model

After defining the high-confidence core and ranking all non-core reactions, our algorithm attempts to sequentially remove each non-core reaction, starting from those ranked at the bottom (lowest evidence). The selected reaction will be removed only if (i) the core set of reaction remains consistent; and (ii) removal does not prevent model from producing any key metabolites. Reactions in high-confidence core set can only be removed when (i) reactions in the negative reaction set (reactions with $\boldsymbol{E_x}(r)$ =0) are needed to enable flux through the high confidence core reactions; (ii) by

removing the high confidence core reactions, more non-core reactions (including those in the negative reaction set) will be removed. Consistency of the core reaction set is confirmed by calculating the maximum and minimum flux for each reaction, and ensuring that at least one is non-zero. As the naïve implementation of flux variability analysis (FVA) is extremely slow, we adapted the *checkModelConsistency* module described by Jerby *et al.* in [135] for optimal performance in Matlab—in particular, we included the option to use the efficient fastFVA algorithm[146].

The list of key metabolites that must be produced from glucose is compiled based on the universal metabolic model validation test in[138]. This includes metabolites in glycolysis, TCA cycle, pentose phosphate pathway, as well as non-essential amino acids, nucleotides, palmital-CoA, cholesterol, and several membrane lipids. Instead of testing the production of all non-essential fatty acids, as in [138], we only tested the production of palmital-CoA, which is derived from palmitate, the first fatty acid produced in fatty acid synthesis, and the precursor of longer chain fatty acids. Similarly, we only tested those membrane lipids that can be derived from glucose and non-essential amino acids. With the addition of essential nutrients like choline, these membrane lipids can be transformed to other membrane lipids such as phosphatidylcholine and sphingomyelin that cannot be directly synthesized from glucose. We only check the production of pyrimidine nucleotides from glucose, as *de novo* pyrimidine synthesis can occur in a variety of tissues [141]. As *de novo* purine synthesis occurs primarily in the liver and other tissues use the salvage pathway [141], we test the ability of all tissues to synthesize purine nucleotides from purines bases and 5-phosphoribosyl 1-pyrophophate (PRPP).

5.7.4 Functional test of liver models

In the amino acid degradation test, the model is only allowed to uptake glucose and the amino acid being tested; all other organic metabolites are constrained to be efflux only. Transport of inorganic compounds (oxygen, carbon dioxide, water, etc.) is unconstrained, except ammonia: as ammonia detoxification is an important hepatic function, only ammonia influx is allowed. The simulation objective function is to maximize the uptake of the amino acid being tested. Using FVA, if the model can allow for finite urea efflux *and* amino acid influx, the amino acid degradation test is declared as passed.

Similarly, in the ammonia detoxification test, only glucose uptake is allowed, and the objective is to maximize ammonia uptake. This test is passed if the model can allow for finite ammonia influx *and* urea efflux. The ethanol detoxification test is the same as ammonia detoxification test, except that no urea efflux is required and ethanol is constrained to be influx.

In the glucogenic test, the model is only allowed to uptake the glucogenic substrate being tested while all other organic compounds, including glucose, are constrained to be efflux only. Ammonia is only allowed to be influx, and urea is only allowed to be efflux. The simulation objective is to maximize glucose secretion. This test is passed if the model can allow for finite glucose efflux. Glucogenic substrates tested are the 18 glucogenic amino acids (all 20 amino acid except leucine and lysine, which are exclusively ketogenic), lactate, pyruvate, and glycerol.

In growth simulation, the widely-used RPMI-1640 tissue culture medium was used. The biomass equation was adopted from [135]. It consists of amino acids, nucleotides, deoxynucleotides, lipids etc. *Recon 1* lacks a reaction accounting for the formation of glycogenin, the primer for glycogen synthesis, so a sink reaction for glycogenin is added to all the liver models to allow for glycogen synthesis.

5.8 Chapter 5 figures and tables

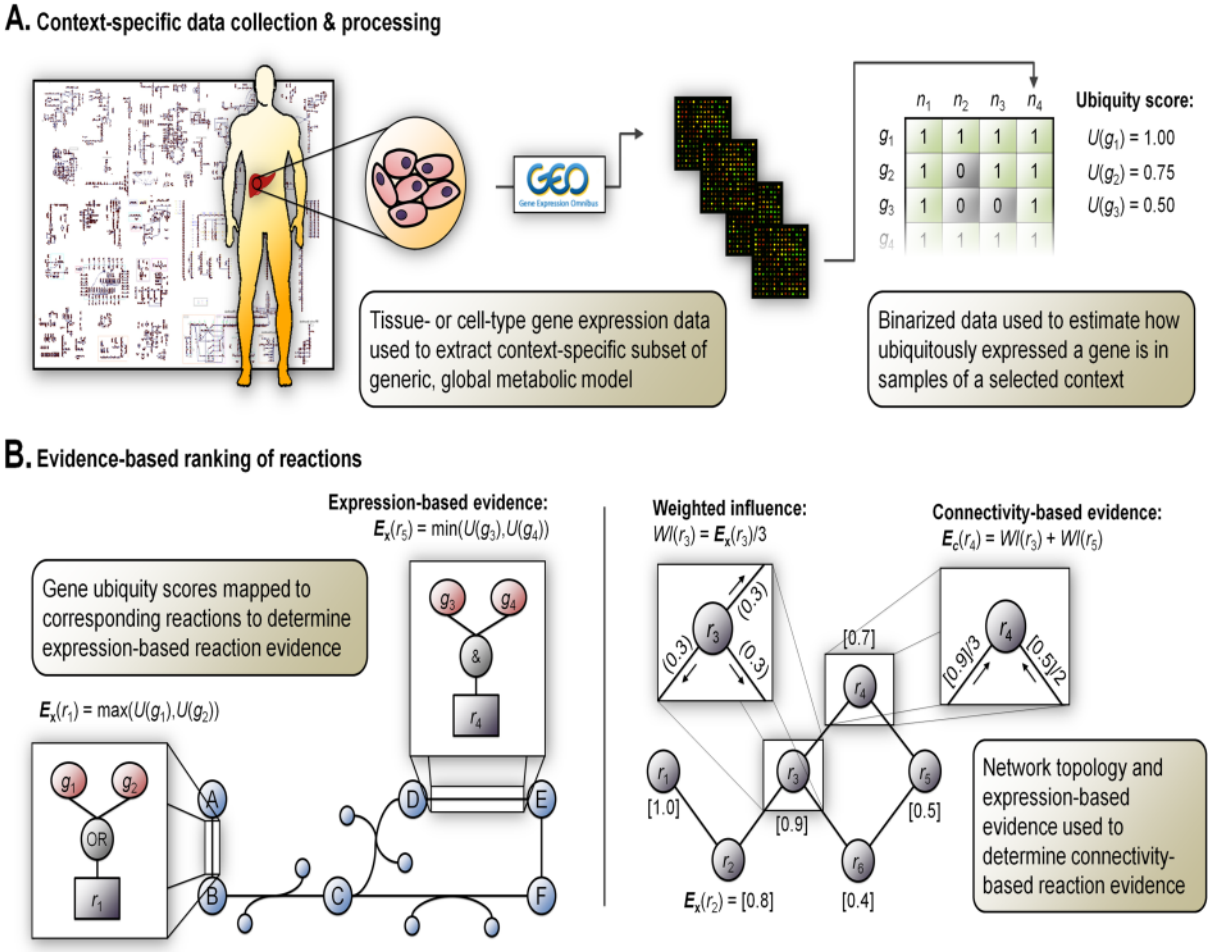

Figure 5.1 mCADRE method overview**.** (**A**) After binarizing context-specific input data, we quantify how often a gene is expressed across samples of the same tissue; this is the ubiquity score $U(g)$ for each gene g. (**B**) From ubiquity scores, we calculate the *expression-based evidence* E_x for each gene-associated reaction. Reactions with sufficiently high E_x are defined as the core reaction set and are included in the tissue-specific model. To deterministically rank non-core reactions with low to moderate expression-based evidence, we introduce a network topology metric to calculate *connectivity-based evidence* E_c.

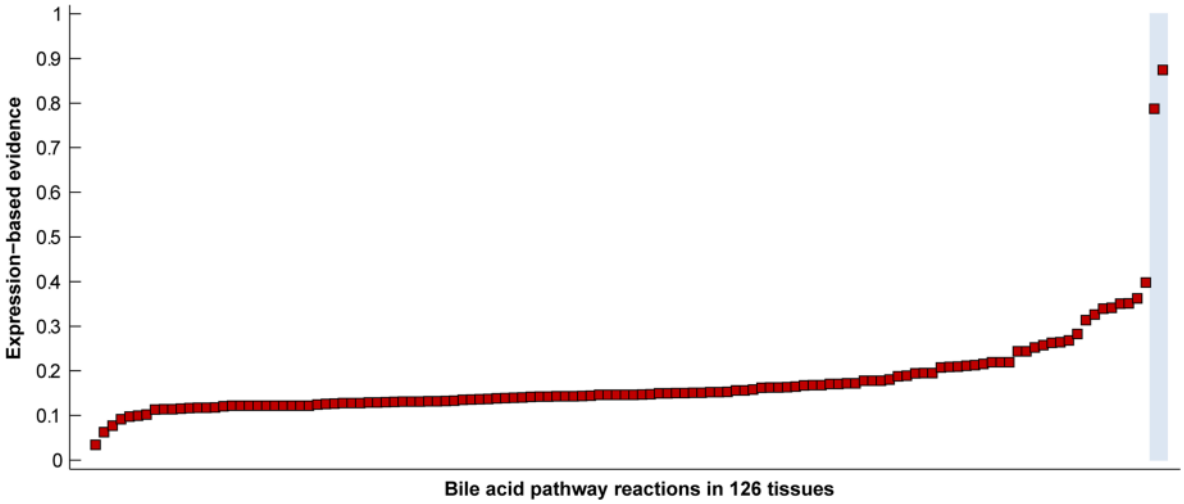

Figure 5.2 **Expression-based evidence for the bile acid synthesis pathway across 126 tissues.** Red squares indicate the average expression-based evidence of all bile acid reactions in the tissue. Normal liver and liver cancer tissues, known to synthesize bile acids, are highlighted in blue.

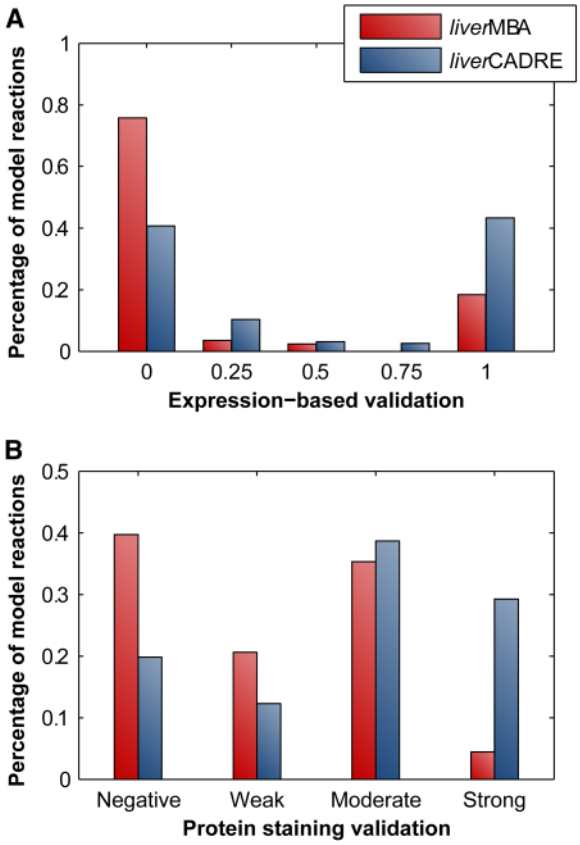

Figure 5.3 **Coverage-based comparison of mCADRE and MBA liver models.** (**A**) Expression-based validation was calculated from an independent microarray data set for reactions unique to *liver*MBA or *liver*CADRE. (**B**) Protein staining data from Human Protein Atlas was mapped to reactions in each model.

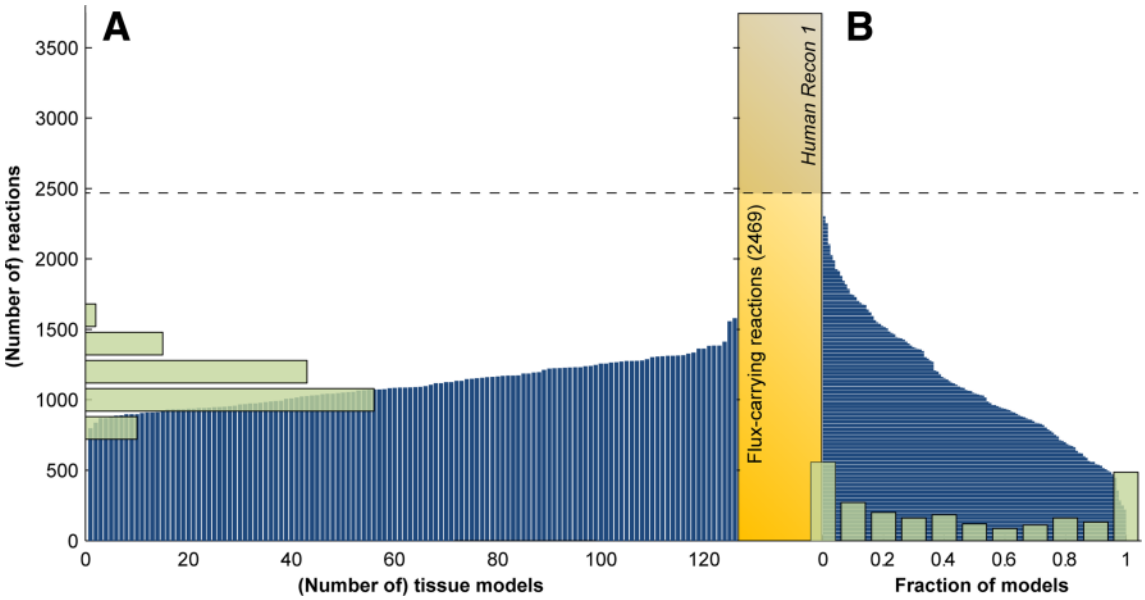

Figure 5.4 **Number and distribution of reactions in TSEM models.** (**A**) Vertical blue lines indicate the number of reactions in each tissue model; green bars show the size distribution across TSEM models. (**B**) Horizontal blue lines indicate the fraction of models in which each *Recon 1* reaction is included; green bars show the frequency distribution across reactions.

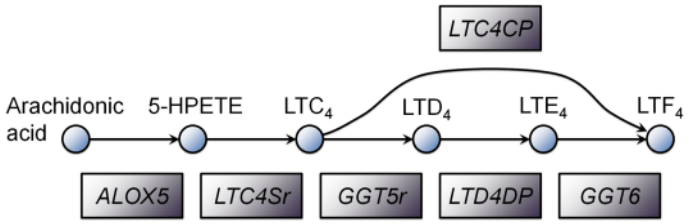

Figure 5.5 **The leukotriene synthesis pathway formed by the reactions occur significantly more often in 17 tumor tissues compared to corresponding normal tissues.** 6 reactions are shown; the other 7 reactions transport metabolites between cellular compartments.

Tuning ratio	Model size	Pos. removed	Neg. removed	%Bile reactions present
0	1194	0.00%	72.63%	70.73%
0.1	1130	1.11%	75.71%	4.88%
0.2	1059	2.77%	78.20%	4.88%
0.25	1037	3.51%	79.03%	4.88%
0.33	1009	5.18%	79.62%	4.88%
0.5	909	11.46%	84.24%	4.88%
1	1216	17.01%	79.50%	9.76%
2	929	31.79%	93.96%	0.00%
3	920	32.72%	94.08%	0.00%
1000	897	34.20%	94.19%	0.00%

Table 5.1. The effect of tuning parameter that balances the inclusion of high confidence positive reactions and the exclusion of high confidence negative reactions.

	***liver*CADRE**	***liver*MBA**
Total reactions	1764	1826
Gene-associated reactions	1192	1167
Total genes	1264	1333
Total metabolites	1401	1360

Table 5.2 . Summary of the mCADRE liver model and the original MBA model.

Functional tests	***liver*CADRE**	***liver*MBA**	***liver*MBA^{Wang}**
gluconeogenesis	13/21	19/21	13/21
Triglycerol synthesis	1/1	0/1	0/1
Amino acid degradation	19/20	20/20	16/20
Ammonia detoxification	1/1	1/1	1/1
Ethanol detoxification	1/1	1/1	1/1
Nucleotide synthesis	8/8	4/8	0/8

Table 5.3 Results of hepatic metabolic function simulations

Pathways[a]	**% complete in brain models**	**% complete non-brain models**	**Rank sum p-value**
Taurine and hypotaurine metabolism	66%	35%	4.14E-08
Fatty acid elongation	67%	37%	1.69E-09
Tyr, Phe, Trp Biosynthesis	77%	56%	4.96E-02
Salvage Pathway	94%	81%	2.26E-02
Cysteine Metabolism	50%	37%	3.61E-03
Alanine and Aspartate Metabolism	73%	61%	2.41E-09
Glutamate metabolism	86%	76%	1.59E-06
Butanoate Metabolism	32%	23%	9.34E-03
Valine, Leucine, and Isoleucine Metabolism	69%	61%	4.00E-02
Transport, Nuclear	33%	25%	2.05E-11

[a]Pathways are sorted by difference in percentage of reactions present in brain vs. non-brain tissues.

Table 5.4 *Recon 1* metabolic pathways differentially represented in brain and non-brain normal tissues.

Pathways[a]	**% complete in tumor models**	**% complete in normal models**	**Rank sum *p*-value**
Folate Metabolism	50%	27%	2.8E-03
Eicosanoid Metabolism	34%	13%	6.6E-04
Fatty acid activation	91%	81%	1.8E-02
Tryptophan metabolism	17%	10%	1.2E-02
Transport, Lysosomal	17%	11%	7.8E-03
Nucleotides	69%	63%	1.9E-04
Aminosugar Metabolism	56%	53%	4.8E-02
Transport, Mitochondrial	25%	23%	3.4E-02
Sphingolipid Metabolism	13%	12%	3.2E-02

[a]Pathways are sorted by difference in percentage of reactions present in tumor vs. normal tissues.

Table 5.5. *Recon 1* metabolic pathways differentially represented in tumor and normal tissues

Reactions[a]	% tumor models	% normal models	Rank sum p-value
ALOX5	72%	6%	8.6E-05
ALOX52	72%	6%	8.6E-05
EX_leuktrC4(e)	72%	6%	8.6E-05
GGT5r	72%	6%	8.6E-05
GGT6	72%	6%	8.6E-05
GLUtr	72%	6%	8.6E-05
GTHRDtr	72%	6%	8.6E-05
LEUKTRA4tr	72%	6%	8.6E-05
LEUKTRC4t	72%	6%	8.6E-05
LEUKTRD4tr	72%	6%	8.6E-05
LTC4CP	72%	6%	8.6E-05
LTC4Sr	72%	6%	8.6E-05
LTD4DP	72%	6%	8.6E-05

[a]These reactions form a pathway that catalyzes the production of leukotrienes from arachidonic acid.

Table 5.6. Top 13 reactions over-represented in tumor tissue models versus corresponding normal tissue models.

	***kidney*TSEM**		***kidney*Reduced**	
	Num.Rxns	**Percentage**	**Num.Rxns**	**Percentage**
Strong	298	37.11%	12	27.91%
Moderate	217	27.02%	19	44.19%
Weak	39	4.86%	1	2.33%
Negative	24	2.99%	2	4.65%
No staining info	225	28.02%	9	20.93%

Table 5.7 Comparion of protein-level evidence of gene-associated reactions unique to each model

Chapter 6. Integrative Reconstruction and Analysis of Genome-Scale Metabolic Models of Commonly Used Breast Cancer Cell Lines

6.1 Metabolic heterogeneity in cancer

Reprogramming of cellular metabolism to facilitate proliferation has been recognized as a hallmark of cancer [178]. The most prominent metabolic feature of proliferating cancer cells is increased consumption of glucose and secretion of lactate even in the presence of oxygen (Warburg effect). However, due to differences in tissue of origin, genetic mutation and expression profile of metabolic genes, cancer cells exhibit great heterogeneity in metabolic reprogramming, leading to a phenomenon dubbed "one hallmark, many faces" [179].

Tissue and cell type are a major contributor to cancer metabolic heterogeneity. Tissue context can affect glycolytic contribution to ATP generation by 2 orders of magnitude, from 0.31% to 64% [180]. A recent study found that Myc-driver liver tumor consumes glutamine, while Myc-driver lung tumor secretes glutamine [29]. Even for cancer of the same tissue of origin, metabolic profile can still differ as a result of genetic mutation or metabolic gene expression. For example, Myc-driver liver tumor consumes glutamine while Met-driver liver tumor secretes glutamine. [29]. Metabolic heterogeneity is particularly prevalent in breast cancer, where multiple clinically relevant subtypes have been established [181]. Luminal subtype of breast cancer cells express glutamine synthetase and are less sensitive to glutamine deprivation than basal subtype of breast cancer cells [182]. Most estrogen receptor-negative breast cancer cells are dependent on the elevated expression of phosphoglycerate dehydrogenase [129]. Loss of the metabolic gene fructose-1,6-bisphosphatase 1 is specifically in basal-like breast cancer [183]. This induces glycolysis and results in increased, macromolecule biosynthesis, and maintenance of ATP production under hypoxia.

Although a majority of research on cancer metabolism focuses on several central metabolic pathways involving glucose and glutamine, other metabolic pathways are increasingly identified as having an important role in proliferation. Recent studies found that the production of N-acetylglucosamine via

hexoamine biosynthesis pathway is required for glycosylation of proteins involved in cancer cell proliferation [184, 185]. S-adenosyl methionine (SAM) is an important co-factor in protein methylation and enzymes involved in SAM synthesis and consumption are found to be critical for maintenance of stem cell state and tumorigenesis [186, 187].

6.2 Integration of multiple types of omic data to reconstruct metabolic models

The great complexity in cancer metabolism, in terms of both the heterogeneity and the multitudes of pathways involved, necessitates the reconstruction of tissue and cell type specific metabolic models that integrates multiple types of functional genomic data to closely represent the specific metabolic profile of the given tissue or cell types. Current models of cancer metabolism are often based on gene microarray data. In this project, I integrated microarray gene expression data, RNA-seq data, proteomic data and metabolite consumption and release data to reconstruct genome-scale metabolic models of 3 cell lines commonly used in breast cancer research; MCF7 is the commonly used estrogen receptor positive and luminal subtype cell line; MDAMB231 is the commonly used estrogen receptor negative and basal substype cell line; MCF10A is the commonly used non-transformed cell line.

4 types of data are used in the reconstruction process that complement each other:

Microarray data. Microarray data for each cell line are collected that come from both unperturbed and perturbed (hypoxia, drug treatment, etc) conditions to capture the wide metabolic capabilities of each cell line. 85, 323, and 129 microarray samples from the Affymetrix U133 Plus 2 platform are collected for MCF10A, MCF7 and MDAMB231 respectively.

RNA-seq data. RNA-seq data was collected for each cell line under normal conditions to capture the baseline metabolic gene expression[188]. RNA-seq data provides a higher resolution than microarray data and also serves as an independent source of evidence about mRNA level evidence.

Proteomic data. For MCF10A and MDAMB231, proteomic data was obtained via iquid chromatography tandem mass spectrometry[189, 190]. For MCF7, antibody staining of metabolic proteins was obtained from the Human Protein Atlas [142].

Metabolite consumption and release (CORE) rates. The above 3 data types provide a static picture of what genes' mRNA and protein are detected in the cell lines. The CORE data set measured cellular consumption and release rates of glucose, different amino acids, nucleotides, fatty acids, and lipids[191], which provides more functional evidence about functional capabilities of the cell line specific metabolic network.

A key step in the reconstruction using mCADRE is to define a high confidence positive core reaction set and a high confidence negative reaction set. A reaction belongs to positive core reaction set if it meets any 2 of the 3 criteria: i) it is detected as "present" by MAS5 algorithm in 25% of microarray samples; ii) it is expressed above 2 FPKM in RNA-seq data; iii) it is detected in proteomic data (2 or more independent peptides in MCF10A and MDAMB231, at least weak staining in MCF7). This ensures that each high confidence core reaction is supported by at least 2 independent data sources. Non-core reactions are iteratively ranked by % of microarray samples they are called "present", expression levels in RNA-seq data, connectivity evidence, and literature evidence, and iteratively removed from the lowest evidence and highest, as described in Chapter 4.

To ensure cell line specific models have the metabolic capabilities measured in the CORE data, the input generic model is fitted to experimentally measured metabolite consumption and release rates. In the pruning process, no reactions are removed if the original optimal fit is affected.

After all above steps, mCADRE produced a model with *reactions* that should be present in each cell line. The gene-reaction association rule was then updated with expression evidence to further increase cell line specificity. For example, the gene-reaction rule for reaction *i* is "gene A or gene B", which means the reaction is present if either A or B is expressed. In a particular cell line, if only gene A is expressed, then the rule needs to be updated to avoid false negatives in *in silico* gene knockout simulation. If reaction *i* is needed for growth, using the original rule, we may come to the false

conclusion that knock down of gene A has no effect. As updating gene-reaction rules affects all subsequent simulations, only the highly confident negative genes (expressed <10% microarray samples AND <0.3 FPKM in RNA-seq data AND not detected in proteomic data) are removed. Note that only genes connected by OR are updated: if the gene-reaction rule is "gene A and gene B" , then it is not updated. More complex gene-reaction rules are iteratively updated from the most basic parts: in " (gene A or gene B) and (gene C or gene D) ", the left and right side of "and" is updated.

After running mCADRE to generate a draft model, I manually curated the cell line specific models to ensure they do not have futile cycles that enable the production of high energy metabolites (ATP, NADH, NADPH, FADH and proton) without any metabolic input.

6.3 Comparison of model prediction and experimental results reveal different types of inconsistency

After manual curation to ensure models pass basic quality checks, *in silico* single gene deletion prediction results were compared with shRNA knockdown data in MCF7 and MDA-MB-231 cell lines [192]. In the experiment, cell lines are infected with shRNA libraries and allowed to grow for 3~4 weeks, and shRNAs that are selectively depleted are chosen as targeting essential genes. Such experiments do not directly measure a phenotypic outcome such as growth rate or ATP level.

	MCF7		**MDA-MB-231**	
	Correct/Total	**Percentage**	**Correct/Total**	**Percentage**
Sensitivity	9/39	0.23	8/42	0.19
Specificity	983/1035	0.95	946/982	0.96
Accuracy	992/1074	0.92	954/1024	0.93

Table 6.1 Comparison of model predicted lethal genes and lethal genes based on shRNA knockdown.

Table 6.1 showed that although 95% experimentally non-lethal genes are corrected predicted as non-lethal (high specificity); only 20% of experimentally lethal genes are correctly predicted as lethal (low sensitivity).

Systematic comparison between model prediction and experimental result reveal four major sources of inconsistency. First, according to model annotation, multiple genes are functionally redundant and any one of them can enable the reaction, but only one gene is experimentally lethal. This is the case even after gene-reaction association rules are updated to include only expressed genes. For example, model predicted that carnitine O-palmitoyltransferase, which transport fatty acid into mitochondria is an essential reaction in MCF7. According to model annotation, any one of the three genes, CPT1A, CPT1B and CPT1C can catalyze the reaction. Both CPT1A and CPT1B are present in the MCF7 cell line, detected at both mRNA and protein level. But only CPT1A is found to be lethal in the shRNA knockdown experiment. Another example is the pantothenate kinase reaction (a key step in Co-enzyme A synthesis), which is predicted to be an essential reaction by the model. All four pantothenate kinase genes (PANK1 to PANK4) are expressed in MCF7 cell line (both mRNA and protein level), but only PANK4 is experimentally lethal. There are three possible explanations for this type of inconsistency. shRNA knockdown screen is known to suffer from off-target effects where shRNAs target unintended genes. Off-target effects may be exacerbated as sequence similarity of functionally redundant genes such as CPT1A and CPT1B is high. Therefore, it is possible that shRNAs targeting CPT1A might also have targeted CPT1B. However, this does not explain why shRNAs targeting CPT1B have no effect on CPT1A. Another possibility is that genes capable of catalyzing the same reaction are not functionally equivalent: the protein encoded by different genes may have different kinetic properties (e.g., affinity for substrates, turn over number), which makes one isozyme plays a more prominent role than others. Unfortunately, current constrained-based modeling approach is not capable of capturing such kinetic differences.

Second, according to model annotation, a reaction is catalyzed by an enzyme complex where multiple genes need to be *all* expressed for the reaction to carry flux, but only one gene is experimentally lethal. For example, the ATP synthase reaction is predicted to be an essential reaction. According to model annotation, for ATP synthase to be active, all its subunits, encoded by different genes, need to be present, and knockdown of any one of the genes will inactivate the enzyme. But only one gene, ATP5E is found to be experimentally lethal. Another example is succinate dehydrogenase, where out of the four subunits (SDHA to SDHD), only SDHC is experimentally lethal. The same is true for NADH dehydrogenase (only NDUFS7 experimentally lethal) and cytochrome oxidase C (only

COX5A experimentally lethal). A possible explanation is the off-target effects of shRNA screen, as discussed above. Another possibility is that the gene-reaction association in model annotation is incorrect.

Third, many experimentally lethal genes are not associated with the synthesis of biomass precursors. To model cancer cell growth, biomass composition typically used in simulation studies often includes amino acids, nucleotides, fatty acids and lipids. However, many experimentally lethal genes have no obvious link to biomass. For example, ATP6V0C is a subunit of vacuolar ATPase that transport proton into lysosome and other organelles (organelle acidification), and this process is essential for protein sorting, zymogen activation, and receptor-mediated endocytosis. None of these processes are reflected in the biomass objective. Even in cancer cells, maximizing growth is not always a primary cellular objective and may fail to capture other necessary processes for proliferation, such as coping with oxidative stress [193]. Even when an enzyme has clear known role in biosynthetic processes, its impact on cell proliferation may not be entirely metabolic: its substrate or product may have functions beyond biosynthetic precursors. For example, mitochondrial fumarate hydratase is a key enzyme in the TCA cycle, and the TCA cycle supplies crucial biosynthetic intermediates such as acetyl-CoA, citrate and a-ketoglutarate. However, fumarate hydratase is actually a tumor suppressor: its inactivation leads to elevated intracellular fumarate, which in turn stabilizes HIF (Hypoxia-inducible factor)[194]. The transcription factor HIF increases the expression of genes in angiogenesis, cell survival, glucose metabolism and invasion. Without experimental knowledge that fumarate can stabilize HIF, metabolic model alone cannot capture all the functional consequences of fumarate hydratase perturbation. This is further complicated by the fact that the same enzyme encoded by the same gene is sometimes distributed in different cellular compartments and performs different functions. Fumatrate hydratase has a mitochondrial form and a cytosolic form, encoded by the same gene. While the mitochondrial form participate in the TCA cycle and inactivation of mitochondrial fumatrate hydratase result in elevated fumarate and stabilization of HIF, the function of the cytosolic form has been elusive. Recently, cytosolic fumarate hydratase and its substrate fumarate are critical compoents of the DNA damage response[195]. Therefore, fumarate hydratase is not only involved in non-metabolic processes, its non-metabolic function is also compartment-specific. In addition to functions in a specific biological process such as HIF stabilization or DNA damage response, a

metabolic gene can also have a broad impact across many processes. For example, Nicotinamide N-methyltransferase (NNMT) consumes methyl units from S-adenosyl methionine to create the stable metabolic product 1-methylnicotinamide, decreasing the methylation potential of cancer cells. As a result, NNMT-expressing cancer cells have an altered epigenetic state that includes hypomethylated histones and other cancer-related proteins combined with heightened expression of protumorigenic gene products[187]. The expression level of NNMT is more than 100 fold higher in the ER-negative MDA-MB-231 cell line than in the ER-positive MCF7 cell line, suggesting certain potential role of NNMT in MDA-MB-231. However, as NNMT decreases methylation potential by creating a stable methyl "sink" as opposed to contribute to biosynthetic precursors, MDA-MB-231 metabolic model did not predict any impact of NNMT on growth.

Fourth, some metabolic genes have important functions beyond metabolism, and therefore are beyond the scope of metabolic models. For example, it has been reported that inhibition of aldolase A, an enzyme in glycolysis, has no effect on glycolytic flux or intracellular ATP level, but disrupted actin-cytoskeleton dynamics and result in increased multi-nucleation. Cell proliferation can be restored by an enzymatically dead aldolase A variant that retain F-actin binding ability, which proved that the lethality of aldolase A knockdown is non-metabolic[196]. Another example of a well-known enzyme having important non-metabolic function is PKM2. Epidermal growth factor receptor (EGFR) activation induces nuclear translocation of PKM2, which binds to phosphorylated β-catenin, leading to histione H3 acetylation and cyclin D1 expression. This PKM2-dependent β-catenin transactivation is instrumental in EGFR-promoted tumour cell proliferation and brain tumour development[197].

6.4 Conclusion

In this project, microarray, RNA-seq, proteomic and metabolic profiling data were integrated to reconstruct genome-scale metabolic models for commonly used breast cancer cell lines. These cell line specific metabolic models are highly confident representations of the metabolic capabilities of the corresponding cell lines: 95% of gene-associated reactions are supported by 2 or more independent sources of data, and are able to consume or release glucose, amino acids, lipids and other metabolites

at experimentally measured rates. These cell line specific models are therefore an important starting point to understand the metabolic states of different breast cancer cell lines.

However, comparison of *in silico* single gene lethality and experimental results revealed many inconsistencies that point to the limitations of current modeling approaches. These inconsistencies are valuable to both improve model annotation and the development of new modeling approaches.

First, metabolic genes may have important characteristics that are not captured by current constrained-based modeling approach. In an isozyme situation (e.g., CPT1A, CPT1B, CPT1C), although each gene product is annotated to be equally capable of catalyzing the same reaction, they may either have different metabolic kinetic properties (e.g., affinity for substrates, turnover number, susceptibility to product inhibition) that are unknown or not incorporated into model annotation. At a even finer level of resolution, it is well known that PKM1 and PKM2, two different splice isoforms of the same PKM gene (pyruvate kinase, muscle), have different kinetic properties: PKM1 has 60% higher kinase activity than PKM2. This difference has profound phenotypic implications: silencing the expression of PKM2 and replacing with PKM1 reverses the Warburg effect (i.e., aerobic glycolysis), a metabolic hallmark of cancer [198]. However, in the commonly used BRENDA enzyme database [199], most information is organized by Enzyme Committee (EC) number. Therefore, CPT1A to CPT1C are annotated to the same EC number; PKLR (pyruvate kinase, liver and RBC) and PKM (including PKM1 and PKM2 isoforms) are also annotated to the same EC number. Future experiments are needed to examine the metabolic differences between different genes encoding different isozymes, and curate a database that records such differences.

Second, it is important to acknowledge that even well studied enzymes (e.g., aldolase A, PKM2) may have non-metabolic functions, and sometimes it is these non-metabolic roles that are more important. Therefore, metabolic model is only part of the toolkit to study biological complexity.

Third, more realistic biological objective functions in addition to maximizing growth are needed to capture the important functions of many metabolic genes. There are several approaches to achieve this. One is to include the production of metabolites with known roles in non-metabolic processes in the

objective function. S-adenosylmethionine (SAM) is an important methyl group donor in many methylation reactions that affect cellular epigenetic state, and the ratio of between SAM and S-adenosylhomocysteine (SAH, product after methyl donation) is a measure of cellular methylation potential. Acetyl-CoA affects histone acetylation and global gene expression. α-ketoglutarate affects DNA methylation. Hexoamines modifies nutrient transporters and growth factor receptors and regulate their translocation. Fatty acids such as myristate, palmitate, farnesyl and geranylgeranyl groups are important substrates in protein modification and affect protein trafficking. Fumarate hydratase stabilizes HIF and increases the expression of pro-survival genes. Including these metabolites in the objective function may enable the identification of reactions that affect cell proliferation via influence on the production and consumption of these metabolites.

Chapter 7. Conclusion

Technical advances have dramatically increased the ability to measure biological systems at increasingly greater levels of resolution, from whole tissues to single cells and subcellular organelle; and more comprehensively from genome, transcriptome, proteome and metabolome. For each biological sample, millions of heterogeneous multi-omic data points are measured. For common diseases, especially cancer, large consortia such as The Cancer Genome Atlas, have gathered multi-omic data from hundreds of patients. The key task is to translation this "big data" into knowledge that help disease diagnosis and increase our understanding of how biological networks operate.

Many statistical approaches have been used to uncover meaningful patterns from large amount of biological data. Statistical tests such as t-test or Wilcoxon ranksum test are used to find genes with different mRNA level between conditions. Logistic regression is used to identify single nucleotide polymorphisms (SNPs) that are associated with phenotypic traits. Supervised approaches such as support vector machines (SVM), k-nearest neighbors (kNN) are used to find accurate classifiers that distinguish phenotypes. Unsupervised approaches such as clustering are used to uncover clinically meaningful subtypes of diseases previously regarded as homogeneous. These tools are unbiased, as they make no *a priori* assumption about what biological features are more important than others. In addition, they often assume different biological features are independent. These statistical approaches are important tools when we know little about the given biological problem, or when the primary objective is to find the most informative features.

However, prior knowledge is a powerful resource to both account for the relations between biological features (pathways, protein-protein interactions, transcription factor-target genes) and reduce the search space to biological features that have known biological functions. Successful incorporation of prior knowledge include gene set enrichment analysis (GSEA) that identify coordinated expression change in *a priori* defined pathways, or pathway-level genome wide association study (pathway GWAS) that identify a group of functionally related genes jointly associated with a trait of interest.

The central theme of my thesis research was to use prior knowledge as biological constraints in the analysis of disease omic data (Figure 7.1). In Chapter 3, interacting Top Scoring Pairs (iTSP) improves upon previous methods by restricting searches to gene pairs with known functional relations. In Chapter 4.2 and 4.3, a more mechanistic constraint is applied to identify interesting expression patterns of reaction pairs that consume the same metabolite. Unlike protein-protein interaction networks (used by iTSP in Chapter 3) that suffer from high false positives, or signaling networks that are very incomplete, metabolic networks are arguably the most well characterized biological networks. This enables more mechanistic interpretation of omic data analysis. For example, if the protein products of two genes interact, relative expression changes between the two genes may not always have straightforward interpretation. On the other hand, if there is a relative expression change between two reactions consuming the same metabolite, it is reasonable to assume that this change may affect the allocation of the given metabolite to different utilization pathways.

Many existing methods test the enrichment of differentially expressed genes in a priori defined gene sets: GSEA typically use biological pathways, while reporter metabolite analysis use genes linked to the same metabolite. Although they facilitate the interpretation of statistical analysis results, they do not account for the mechanistic link between biological features. For example, expression aberration or deleterious mutation in any single subunit of an enzyme complex is expected to have similar phenotypic consequences. Similarly, such aberrations along a contiguous chain of metabolic reactions are also expected to have similar effects: SNPs in different genes encoding enzymes in heme biosynthesis pathways all result in porphyria. Based on such mechanistic links, in a hypothetical scenario to "rediscover" causal pathways linked to porphyria, heme biosynthesis pathway can be identified without resorting to statistical enrichment. To exploit such mechanistic links on a network level, in Chapter 4.4, a new analysis approach identified the preferential allocation of metabolites at metabolic branch points by considering the constraints imposed by the gene expression states of all reactions in the network, not just the reaction pairs at the branch point. This analysis is an important step in the development of "mechanistic classifiers", where different phenotypes are distinguished by their network states. The network state is constrained by both the expression, phosphorylation or methylation status its components and the mechanistic links between its components based on prior knowledge.

The other theme of my research is the contextualization and refinement of prior knowledge by omic data (Figure 7.2). The generic human metabolic network is reconstructed from extensive literature curation and represents the union of known metabolic capabilities of all human tissues and cell types. In Chapter 5, I developed mCADRE to contextualize the generic human metabolic network using transcriptomic, proteomic and metabolomic data of a given tissue. These tissue- and cell-type-specific models can then be compared with context-specific perturbation data (e.g. shRNA knockdown, enzyme inhibition) to reveal knowledge gaps in our current understanding of human metabolism.

Taken together, the main achievement of my dissertation is the development of various computational pipelines that integrates statistical and mechanistic modeling approaches. These comoputational methods incorporate prior knowledge into disease omic data analysis and contextualize prior knowledge with omic data. Large consortia continue to generate and process large omic data sets (e.g. The Cancer Genome Atlas), as well as to curate and catalogue prior knowledge (e.g., the Human Recon 2). The computational tools developed in my thesis have substantial promise to harness the power of omic data and prior knowledge to generate new biological hypotheses and insights.

7.1 Chapter 7 figures

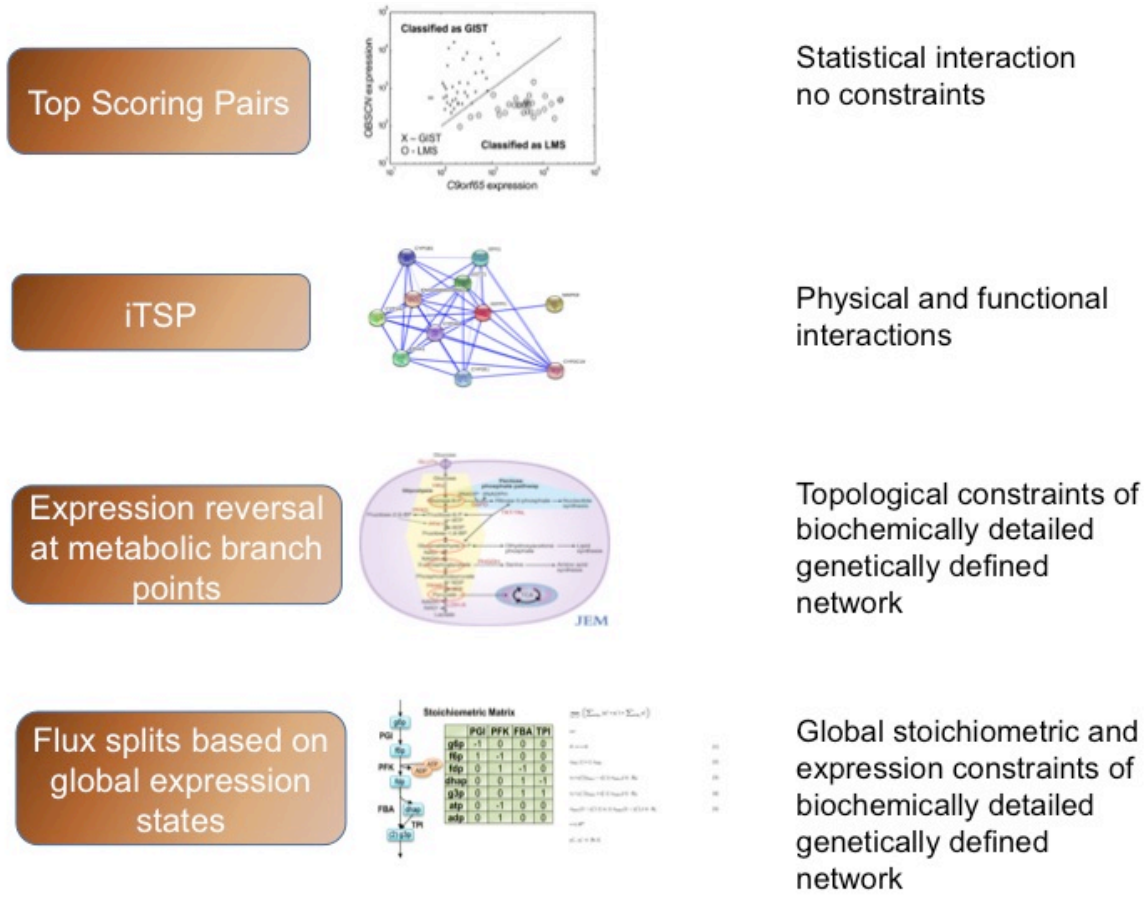

Figure 7.1. Successively more mechanistic biological constraints are applied to disease omic data analysis to improve diagnosis and generate new hypotheses.

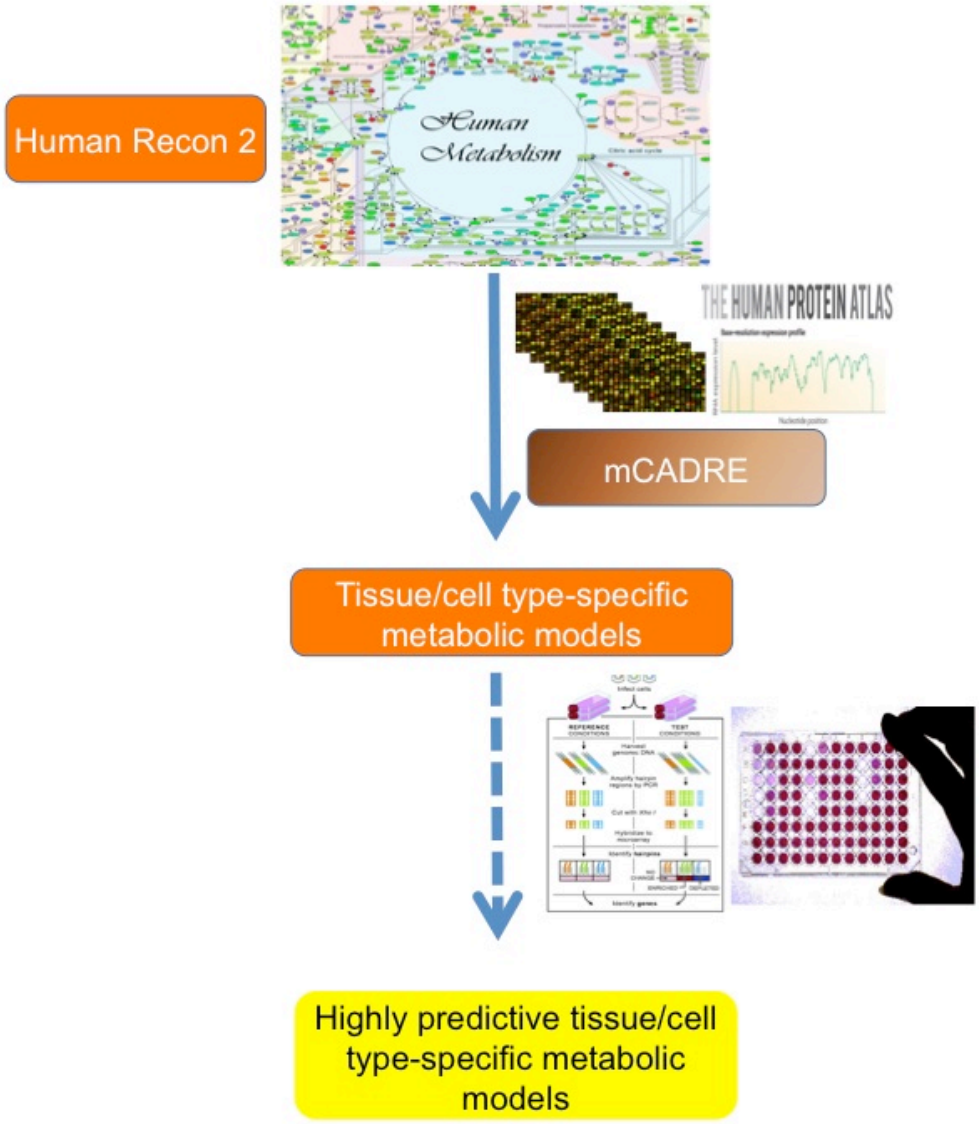

Figure 7.2. Contextualization and refinement of prior knowledge (e.g. metabolic networks) using omic data. Tissue-specific metabolic models are inferred from generic human metabolic network based on various omic data evidence. These models are further refined by comparison with tissue and cell-line-specific perturbation data (e.g., shRNA knockdown).

References

1. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
2. Kamburov, A., et al., *The ConsensusPathDB interaction database: 2013 update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D793-800.
3. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
4. *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
5. *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.
6. *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.
7. *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 2012. **489**(7417): p. 519-25.
8. *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
9. Kandoth, C., et al., *Integrated genomic characterization of endometrial carcinoma.* Nature, 2013. **497**(7447): p. 67-73.
10. Reinhold, W.C., et al., *CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set.* Cancer Res, 2012. **72**(14): p. 3499-511.
11. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
12. Abaan, O.D., et al., *The Exomes of the NCI-60 Panel: A Genomic Resource for Cancer Biology and Systems Pharmacology.* Cancer Res, 2013. **73**(14): p. 4372-82.
13. Sung, J., et al., *Molecular signatures from omics data: from chaos to consensus.* Biotechnol J, 2012. **7**(8): p. 946-57.
14. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.
15. Paik, S., et al., *A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.* N Engl J Med, 2004. **351**(27): p. 2817-26.
16. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
17. Tibshirani, R., et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* Proc Natl Acad Sci U S A, 2002. **99**(10): p. 6567-72.
18. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
19. Draghici, S., et al., *A systems biology approach for pathway level analysis.* Genome Res, 2007. **17**(10): p. 1537-45.
20. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis.* Mol Syst Biol, 2007. **3**: p. 140.
21. Duarte, N.C., et al., *Global reconstruction of the human metabolic network based on genomic and bibliomic data.* Proc Natl Acad Sci U S A, 2007. **104**(6): p. 1777-82.
22. Thiele, I., et al., *A community-driven global reconstruction of human metabolism.* Nat Biotechnol, 2013. **31**(5): p. 419-25.
23. Lewis, N.E., H. Nagarajan, and B.O. Palsson, *Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods.* Nat Rev Microbiol, 2012. **10**(4): p. 291-305.
24. Jerby, L., T. Shlomi, and E. Ruppin, *Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism.* Mol Syst Biol, 2010. **6**: p. 401.

25. Agren, R., et al., *Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT.* PLoS Comput Biol, 2012. **8**(5): p. e1002518.
26. Wang, Y., J.A. Eddy, and N.D. Price, *Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE.* BMC Syst Biol, 2012. **6**: p. 153.
27. Cairns, R.A., I.S. Harris, and T.W. Mak, *Regulation of cancer cell metabolism.* Nat Rev Cancer, 2011. **11**(2): p. 85-95.
28. Oermann, E.K., et al., *Alterations of metabolic genes and metabolites in cancer.* Seminars in Cell & Developmental Biology, 2012. **23**(4): p. 370-380.
29. Yuneva, Mariia O., et al., *The Metabolic Profile of Tumors Depends on Both the Responsible Genetic Lesion and Tissue Type.* Cell metabolism, 2012. **15**(2): p. 157-170.
30. Hu, J., et al., *Heterogeneity of tumor-induced gene expression changes in the human metabolic network.* Nat Biotech, 2013. **31**(6): p. 522-529.
31. Frezza, C., et al., *Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase.* Nature, 2011. **477**(7363): p. 225-228.
32. Ramaswamy, S., et al., *A molecular signature of metastasis in primary solid tumors.* Nat Genet, 2003. **33**(1): p. 49-54.
33. Gomez Ravetti, M. and P. Moscato, *Identification of a 5-protein biomarker molecular signature for predicting Alzheimer's disease.* PloS one, 2008. **3**(9): p. e3111.
34. Price, N.D., et al., *Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(9): p. 3414-3419.
35. Schadt, E.E., *Molecular networks as sensors and drivers of common human diseases.* Nature, 2009. **461**(7261): p. 218-23.
36. Zender, L., et al., *Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach.* Cell, 2006. **125**(7): p. 1253-67.
37. Varambally, S., et al., *Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression.* Cancer cell, 2005. **8**(5): p. 393-406.
38. Hood, L., et al., *Systems biology and new technologies enable predictive and preventative medicine.* Science, 2004. **306**(5696): p. 640-3.
39. Mehrabian, M., et al., *Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits.* Nat Genet, 2005. **37**(11): p. 1224-33.
40. Pericak-Vance, M.A., et al., *Complete genomic screen in late-onset familial Alzheimer disease. Evidence for a new locus on chromosome 12.* JAMA : the journal of the American Medical Association, 1997. **278**(15): p. 1237-41.
41. Hur, J., et al., *The identification of gene expression profiles associated with progression of human diabetic neuropathy.* Brain : a journal of neurology, 2011. **134**(Pt 11): p. 3222-35.
42. Friedman, D.R., et al., *A genomic approach to improve prognosis and predict therapeutic response in chronic lymphocytic leukemia.* Clinical cancer research : an official journal of the American Association for Cancer Research, 2009. **15**(22): p. 6947-55.
43. Cohen, A.L., et al., *A pharmacogenomic method for individualized prediction of drug sensitivity.* Molecular systems biology, 2011. **7**: p. 513.
44. Xie, L., et al., *Novel computational approaches to polypharmacology as a means to define responses to individual drugs.* Annu Rev Pharmacol Toxicol, 2012. **52**: p. 361-79.
45. Hines, A., et al., *Discovery of metabolic signatures for predicting whole organism toxicology.* Toxicol Sci, 2010. **115**(2): p. 369-78.

46. Guerreiro, N., et al., *Toxicogenomics in drug development.* Toxicol Pathol, 2003. **31**(5): p. 471-9.
47. Bovelstad, H.M., et al., *Predicting survival from microarray data--a comparative study.* Bioinformatics, 2007. **23**(16): p. 2080-7.
48. Pittman, J., et al., *Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(22): p. 8431-6.
49. Buyse, M., et al., *Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer.* Journal of the National Cancer Institute, 2006. **98**(17): p. 1183-92.
50. van de Vijver, M.J., et al., *A gene-expression signature as a predictor of survival in breast cancer.* The New England journal of medicine, 2002. **347**(25): p. 1999-2009.
51. Ntzani, E.E. and J.P. Ioannidis, *Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment.* Lancet, 2003. **362**(9394): p. 1439-44.
52. Feng, Z., R. Prentice, and S. Srivastava, *Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective.* Pharmacogenomics, 2004. **5**(6): p. 709-19.
53. Brenner, D.E. and D.P. Normolle, *Biomarkers for cancer risk, early detection, and prognosis: the validation conundrum.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2007. **16**(10): p. 1918-20.
54. Ransohoff, D.F., *Bias as a threat to the validity of cancer molecular-marker research.* Nature reviews. Cancer, 2005. **5**(2): p. 142-9.
55. McIntosh, M., et al., *Ovarian cancer early detection claims are biased.* Clinical cancer research : an official journal of the American Association for Cancer Research, 2008. **14**(22): p. 7574; author reply 7577-9.
56. Hughes, V., *Markers of dispute.* Nature medicine, 2009. **15**(12): p. 1339-43.
57. Simon, R., *Roadmap for developing and validating therapeutically relevant genomic classifiers.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2005. **23**(29): p. 7332-41.
58. McDonnell, B., et al., *Cardiac biomarkers and the case for point-of-care testing.* Clin Biochem, 2009. **42**(7-8): p. 549-61.
59. Ideker, T., J. Dutkowski, and L. Hood, *Boosting signal-to-noise in complex biology: prior knowledge is power.* Cell, 2011. **144**(6): p. 860-3.
60. Dougherty, E.R., *Small sample issues for microarray-based classification.* Comp Funct Genomics, 2001. **2**(1): p. 28-34.
61. Orešič, M., Hyötyläinen, T., Herukka, S.-K., Sysi-Aho, M., Mattila, I., Seppänan-Laakso, T., Julkunen, P., Gopalacharyulu V., Hallikainen M., Koikkalainen J., Kivipelto M., Helisalmi S., Lötjönen, J. and Soininen, H., *Metabolome in progression to Alzheimer's disease.* Translational Psychiatry, 2011. **1**(e57): p. 2158-3188.
62. Barba, I., et al., *Alzheimer's disease beyond the genomic era: nuclear magnetic resonance (NMR) spectroscopy-based metabolomics.* J Cell Mol Med, 2008. **12**(5A): p. 1477-85.
63. Greenberg, N., et al., *A proposed metabolic strategy for monitoring disease progression in Alzheimer's disease.* Electrophoresis, 2009. **30**(7): p. 1235-9.
64. Parkinson, H., et al., *ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments.* Nucleic acids research, 2011. **39**(Database issue): p. D1002-4.
65. Kodama, Y., M. Shumway, and R. Leinonen, *The sequence read archive: explosive growth of sequencing data.* Nucleic acids research, 2011.
66. Akey, J.M., et al., *On the design and analysis of gene expression studies in human populations.* Nat Genet, 2007. **39**(7): p. 807-8; author reply 808-9.

67. Allison, D.B., et al., *Microarray data analysis: from disarray to consolidation and consensus.* Nature reviews. Genetics, 2006. **7**(1): p. 55-65.
68. Scherer, A., *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Probability and Statistics. 2009: Wiley. 272.
69. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nature reviews. Genetics, 2010. **11**(10): p. 733-9.
70. Leek, J.T. and J.D. Storey, *Capturing heterogeneity in gene expression studies by surrogate variable analysis.* PLoS genetics, 2007. **3**(9): p. 1724-35.
71. Reimers, M., *Making informed choices about microarray data analysis.* PLoS computational biology, 2010. **6**(5): p. e1000786.
72. Braga-Neto, U.M. and E.R. Dougherty, *Is cross-validation valid for small-sample microarray classification?* Bioinformatics, 2004. **20**(3): p. 374-380.
73. Ambroise, C. and G.J. McLachlan, *Selection bias in gene extraction on the basis of microarray gene-expression data.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(10): p. 6562-6.
74. Baggerly, K., *Disclose all data in publications.* Nature, 2010. **467**(7314): p. 401.
75. John P. A. Ioannidis, M.J.K., *Improving Validation Practices in "Omics" Research.* Science, 2011. **334**(6060): p. 1230-1232.
76. Dudley, J.T., et al., *Disease signatures are robust across tissues and experiments.* Molecular systems biology, 2009. **5**: p. 307.
77. Miller, J.A., S. Horvath, and D.H. Geschwind, *Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(28): p. 12698-12703.
78. Xu, L., et al., *Merging microarray data from separate breast cancer studies provides a robust prognostic test.* BMC bioinformatics, 2008. **9**: p. 125.
79. Hwang, D., et al., *A data integration methodology for systems biology.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(48): p. 17296-301.
80. Hwang, D., et al., *A data integration methodology for systems biology: experimental verification.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(48): p. 17302-7.
81. Network, T.C.G.A.R., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
82. English, S.B. and A.J. Butte, *Evaluation and integration of 49 genome-wide experiments and the prediction of previously unknown obesity-related genes.* Bioinformatics, 2007. **23**(21): p. 2910-7.
83. Lu, T.P., et al., *Integrated analyses of copy number variations and gene expression in lung adenocarcinoma.* PloS one, 2011. **6**(9): p. e24829.
84. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis.* Molecular systems biology, 2007. **3**: p. 140.
85. Chandrasekaran, S. and N.D. Price, *Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(41): p. 17845-50.
86. Hwang, D., et al., *A systems approach to prion disease.* Molecular systems biology, 2009. **5**: p. 252.
87. Slater, T., C. Bouton, and E.S. Huang, *Beyond data integration.* Drug discovery today, 2008. **13**(13-14): p. 584-9.
88. Li, C. and H. Li, *Network-constrained regularization and variable selection for analysis of genomic data.* Bioinformatics, 2008. **24**(9): p. 1175-82.

89. Witten, D.M. and R. Tibshirani, *Covariance-regularized regression and classification for high-dimensional problems.* J R Stat Soc Series B Stat Methodol, 2009. **71**(3): p. 615-636.
90. Caiyan Li, H.L., *Variable selection and regression analysis for graph-structured covariates with an application to genomics.* Annals of Applied Statistics, 2010. **4**(3): p. 1498-1516.
91. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-50.
92. Eddy, J.A., et al., *Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC).* PLoS computational biology, 2010. **6**(5): p. e1000792.
93. Nibbe, R.K., M. Koyuturk, and M.R. Chance, *An integrative -omics approach to identify functional sub-networks in human colorectal cancer.* PLoS computational biology, 2010. **6**(1): p. e1000639.
94. Irish, J.M., et al., *Single cell profiling of potentiated phospho-protein networks in cancer cells.* Cell, 2004. **118**(2): p. 217-28.
95. Hale, M.B., et al., *Stage dependent aberrant regulation of cytokine-STAT signaling in murine systemic lupus erythematosus.* PloS one, 2009. **4**(8): p. e6756.
96. Ioannidis, J.P., *Why most published research findings are false.* PLoS Med, 2005. **2**(8): p. e124.
97. Geman, D., et al., *Classifying gene expression profiles from pairwise mRNA comparisons.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article19.
98. Tan, A.C., et al., *Simple decision rules for classifying human cancers from gene expression profiles.* Bioinformatics, 2005. **21**(20): p. 3896-904.
99. Lin, X., et al., *The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations.* BMC Bioinformatics, 2009. **10**: p. 256.
100. Magis, A.T. and N.D. Price, *The top-scoring 'N' algorithm: a generalized relative expression classification method from small numbers of biomolecules.* BMC Bioinformatics, 2012. **13**: p. 227.
101. Ein-Dor, L., et al., *Outcome signature genes in breast cancer: is there a unique set?* Bioinformatics, 2005. **21**(2): p. 171-178.
102. Ein-Dor, L., O. Zuk, and E. Domany, *Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(15): p. 5923-5928.
103. Franceschini, A., et al., *STRING v9.1: protein-protein interaction networks, with increased coverage and integration.* Nucleic Acids Research, 2013. **41**(Database issue): p. D808-15.
104. *The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.* Nat Biotech, 2010. **28**(8): p. 827-838.
105. Desta, Z., et al., *Clinical significance of the cytochrome P450 2C19 genetic polymorphism.* Clin Pharmacokinet, 2002. **41**(12): p. 913-58.
106. Chau, T.K., et al., *Genotype analysis of the CYP2C19 gene in HCV-seropositive patients with cirrhosis and hepatocellular carcinoma.* Life Sci, 2000. **67**(14): p. 1719-24.
107. Burim, R.V., et al., *Polymorphisms in glutathione S-transferases GSTM1, GSTT1 and GSTP1 and cytochromes P450 CYP2E1 and CYP1A1 and susceptibility to cirrhosis or pancreatitis in alcoholics.* Mutagenesis, 2004. **19**(4): p. 291-8.
108. Ghobadloo, S.M., et al., *GSTP1, GSTM1, and GSTT1 genetic polymorphisms in patients with cryptogenic liver cirrhosis.* J Gastrointest Surg, 2004. **8**(4): p. 423-7.
109. Gilman, S.R., et al., *Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses.* Neuron, 2011. **70**(5): p. 898-907.

110. Ideker, T. and R. Sharan, *Protein networks in disease.* Genome Research, 2008. **18**(4): p. 644-652.
111. Schaefer, M.H., et al., *HIPPIE: Integrating protein interaction networks with experiment based quality scores.* Plos One, 2012. **7**(2): p. e31826.
112. Earls, J.C., et al., *AUREA: an open-source software system for accurate and user-friendly identification of relative expression molecular signatures.* BMC Bioinformatics, 2013. **14**: p. 78.
113. Mas, V.R., et al., *Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma.* Mol Med, 2009. **15**(3-4): p. 85-94.
114. Farci, P., et al., *B cell gene signature with massive intrahepatic production of antibodies to hepatitis B core antigen in hepatitis B virus–associated acute liver failure.* Proceedings of the National Academy of Sciences, 2010. **107**(19): p. 8766-8771.
115. Archer, K.J., et al., *Identifying genes for establishing a multigenic test for hepatocellular carcinoma surveillance in hepatitis C virus-positive cirrhotic patients.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(11): p. 2929-32.
116. Wurmbach, E., et al., *Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma.* Hepatology, 2007. **45**(4): p. 938-47.
117. Chiang, D.Y., et al., *Focal gains of VEGFA and molecular classification of hepatocellular carcinoma.* Cancer Res, 2008. **68**(16): p. 6779-88.
118. Sarasin-Filipowicz, M., et al., *Interferon signaling and treatment outcome in chronic hepatitis C.* Proc Natl Acad Sci U S A, 2008. **105**(19): p. 7034-9.
119. Honda, M., et al., *Differential interferon signaling in liver lobule and portal area cells under treatment for chronic hepatitis C.* J Hepatol, 2010. **53**(5): p. 817-26.
120. Wu, Z.J., et al., *A model-based background adjustment for oligonucleotide expression arrays.* Journal of the American Statistical Association, 2004. **99**(468): p. 909-917.
121. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.
122. Magis, A. and N. Price, *The top-scoring 'N' algorithm: a generalized relative expression classification method from small numbers of biomolecules.* BMC Bioinformatics, 2012. **13**(1): p. 227.
123. Duarte, N.C., et al., *Global reconstruction of the human metabolic network based on genomic and bibliomic data.* Proceedings of the National Academy of Sciences, 2007. **104**(6): p. 1777.
124. Ma, H., et al., *The Edinburgh human metabolic network reconstruction and its functional analysis.* Molecular Systems Biology, 2007. **3**: p. 135.
125. Bordbar, A. and B.O. Palsson, *Using the reconstructed genome-scale human metabolic network to study physiology and pathology.* Journal of Internal Medicine, 2012. **271**(2): p. 131-141.
126. Lazar, M.A. and M.J. Birnbaum, *Physiology. De-meaning of metabolism.* Science, 2012. **336**(6089): p. 1651-2.
127. Ramsköld, D., et al., *An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data.* PLoS Comput Biol, 2009. **5**(12): p. e1000598.
128. Nilsson, L.M., et al., *Mouse genetics suggests cell-context dependency for Myc-regulated metabolic enzymes during tumorigenesis.* PLoS Genet, 2012. **8**(3): p. e1002573.
129. Possemato, R., et al., *Functional genomics reveal that the serine synthesis pathway is essential in breast cancer.* Nature, 2011. **476**(7360): p. 346-350.
130. Locasale, J.W., et al., *Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis.* Nat Genet, 2011. **43**(9): p. 869-74.
131. Shlomi, T., et al., *Network-based prediction of human tissue-specific metabolism.* Nature Biotechnology, 2008. **26**(9): p. 1003-10.

132. Bordbar, A., et al., *Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions.* Mol Syst Biol, 2010. **6**.
133. Chang, R.L., et al., *Drug Off-Target Effects Predicted Using Structural Analysis in the Context of a Metabolic Network Model.* PLoS Comput Biol, 2010. **6**(9): p. e1000938.
134. Gille, C., et al., *HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology.* Mol Syst Biol, 2010. **6**.
135. Jerby, L., T. Shlomi, and E. Ruppin, *Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism.* Mol Syst Biol, 2010. **6**.
136. Folger, O., et al., *Predicting selective drug targets in cancer through metabolic networks.* Mol Syst Biol, 2011. **7**.
137. Becker, S.A. and B.O. Palsson, *Context-specific metabolic networks are consistent with experiments.* PLoS Computational Biology, 2008. **4**(5): p. e1000082.
138. Bordbar, A., et al., *A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology.* BMC Systems Biology, 2011. **5**: p. 180.
139. Guillermet-Guibert, J., et al., *Targeting the sphingolipid metabolism to defeat pancreatic cancer cell resistance to the chemotherapeutic gemcitabine drug.* Mol Cancer Ther, 2009. **8**(4): p. 809-20.
140. McCall, M.N., et al., *The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes.* Nucleic Acids Research, 2011. **39**(suppl 1): p. D1011-D1015.
141. Rosenthal, M.D. and R.H. Glew, *Medical biochemistry : human metabolism in health and disease*. 2009, Oxford: Wiley & Sons.
142. Uhlen, M., et al., *Towards a knowledge-based Human Protein Atlas.* Nat Biotechnol, 2010. **28**(12): p. 1248-50.
143. Ohno, S. and S. Nakajin, *Determination of mRNA expression of human UDP-glucuronosyltransferases and application for localization in various human tissues by real-time reverse transcriptase-polymerase chain reaction.* Drug Metab Dispos, 2009. **37**(1): p. 32-40.
144. Shelby, M.K., et al., *Tissue mRNA expression of the rat UDP-glucuronosyltransferase gene family.* Drug Metab Dispos, 2003. **31**(3): p. 326-33.
145. Mahadevan, R. and C.H. Schilling, *The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.* Metab Eng, 2003. **5**(4): p. 264-76.
146. Gudmundsson, S. and I. Thiele, *Computationally efficient flux variability analysis.* BMC Bioinformatics, 2010. **11**: p. 489.
147. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature Reviews Genetics, 2009. **10**(1): p. 57-63.
148. Krupp, M., et al., *RNA-Seq Atlas – A reference database for gene expression profiling in normal tissue by next generation sequencing.* Bioinformatics, 2012.
149. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-342.
150. Uhlen, M., et al., *Towards a knowledge-based Human Protein Atlas.* Nat Biotech, 2010. **28**(12): p. 1248-1250.
151. Moore, S.A., *Polyunsaturated fatty acid synthesis and release by brain-derived cells in vitro.* J Mol Neurosci, 2001. **16**(2-3): p. 195-200; discussion 215-21.
152. Moore, S.A., et al., *Astrocytes, Not Neurons, Produce Docosahexaenoic Acid (22:6ω-3) and Arachidonic Acid (20:4ω-6).* Journal of Neurochemistry, 1991. **56**(2): p. 518-524.
153. Kuhajda, F.P., *Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology.* Nutrition, 2000. **16**(3): p. 202-8.
154. Menendez, J.A. and R. Lupu, *Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis.* Nat Rev Cancer, 2007. **7**(10): p. 763-777.

155. Mashima, T., H. Seimiya, and T. Tsuruo, *De novo fatty-acid synthesis and related pathways as molecular targets for cancer therapy.* Br J Cancer, 2009. **100**(9): p. 1369-72.

156. Wang, D. and R.N. Dubois, *Eicosanoids and cancer.* Nat Rev Cancer, 2010. **10**(3): p. 181-93.

157. Ogretmen, B. and Y.A. Hannun, *Biologically active sphingolipids in cancer pathogenesis and treatment.* Nat Rev Cancer, 2004. **4**(8): p. 604-616.

158. Ye, Y.N., et al., *A mechanistic study of colon cancer growth promoted by cigarette smoke extract.* Eur J Pharmacol, 2005. **519**(1-2): p. 52-7.

159. Cianchi, F., et al., *Inhibition of 5-lipoxygenase by MK886 augments the antitumor activity of celecoxib in human colon cancer cells.* Mol Cancer Ther, 2006. **5**(11): p. 2716-26.

160. Peters-Golden, M. and W.R. Henderson, Jr., *Leukotrienes.* N Engl J Med, 2007. **357**(18): p. 1841-54.

161. Tsopanoglou, N.E., E. Pipili-Synetos, and M.E. Maragoudakis, *Leukotrienes C4 and D4 promote angiogenesis via a receptor-mediated interaction.* Eur J Pharmacol, 1994. **258**(1-2): p. 151-4.

162. Paruchuri, S., et al., *The pro-inflammatory mediator leukotriene D4 induces phosphatidylinositol 3-kinase and Rac-dependent migration of intestinal epithelial cells.* J Biol Chem, 2005. **280**(14): p. 13538-44.

163. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets—10 years on.* Nucleic Acids Research, 2011. **39**(suppl 1): p. D1005-D1010.

164. Davis, S. and P.S. Meltzer, *GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor.* Bioinformatics, 2007. **23**(14): p. 1846-1847.

165. Dudley, J. and A.J. Butte, *Enabling integrative genomic analysis of high-impact human diseases through text mining.* Pac Symp Biocomput, 2008: p. 580-91.

166. Desvergne, B., L. Michalik, and W. Wahli, *Transcriptional Regulation of Metabolism.* Physiol. Rev., 2006. **86**(2): p. 465-514.

167. Fajans, S.S., G.I. Bell, and K.S. Polonsky, *Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young.* N Engl J Med, 2001. **345**(13): p. 971-80.

168. Evans, R.M., G.D. Barish, and Y.X. Wang, *PPARs and the complex journey to obesity.* Nature Medicine, 2004. **10**(4): p. 355-361.

169. Maeda, K., et al., *Adipocyte/macrophage fatty acid binding proteins control integrated metabolic responses in obesity and diabetes.* Cell Metabolism, 2005. **1**(2): p. 107-119.

170. Furuhashi, M., et al., *Adipocyte/macrophage fatty acid–binding proteins contribute to metabolic deterioration through actions in both macrophages and adipocytes in mice.* The Journal of Clinical Investigation, 2008. **118**(7): p. 2640-2650.

171. Allaman, I., M. Belanger, and P.J. Magistretti, *Astrocyte-neuron metabolic relationships: for better and for worse.* Trends Neurosci, 2011. **34**(2): p. 76-87.

172. Schellenberger, J., et al., *BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions.* BMC Bioinformatics, 2010. **11**(1): p. 213.

173. Schellenberger, J., et al., *Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0.* Nat. Protocols, 2011. **6**(9): p. 1290-1307.

174. Marioni, J.C., et al., *RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays.* Genome Research, 2008. **18**(9): p. 1509-1517.

175. Roth, R., et al., *Gene expression analyses reveal molecular relationships among 20 regions of the human CNS.* Neurogenetics, 2006. **7**(2): p. 67-80.

176. Liao, Y.L., et al., *Identification of SOX4 target genes using phylogenetic footprinting-based prediction from expression microarrays suggests that overexpression of SOX4 potentiates metastasis in hepatocellular carcinoma.* Oncogene, 2008. **27**(42): p. 5578-5589.

177. Hubbell, E., W.M. Liu, and R. Mei, *Robust estimators for expression analysis.* Bioinformatics, 2002. **18**(12): p. 1585-92.

178. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

179. Cantor, J.R. and D.M. Sabatini, *Cancer cell metabolism: one hallmark, many faces.* Cancer Discov, 2012. **2**(10): p. 881-98.

180. Zu, X.L. and M. Guppy, *Cancer metabolism: facts, fantasy, and fiction.* Biochem Biophys Res Commun, 2004. **313**(3): p. 459-65.

181. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. **27**(8): p. 1160-7.

182. Kung, H.-N., J.R. Marks, and J.-T. Chi, *Glutamine Synthetase Is a Genetic Determinant of Cell Type–Specific Glutamine Independence in Breast Epithelia.* PLoS Genet, 2011. **7**(8): p. e1002229.

183. Dong, C., et al., *Loss of FBP1 by Snail-mediated repression provides metabolic advantages in basal-like breast cancer.* Cancer Cell, 2013. **23**(3): p. 316-31.

184. Wellen, K.E., et al., *The hexosamine biosynthetic pathway couples growth factor-induced glutamine uptake to glucose metabolism.* Genes Dev, 2010. **24**(24): p. 2784-99.

185. Itkonen, H.M., et al., *O-GlcNAc transferase integrates metabolic pathways to regulate the stability of c-MYC in human prostate cancer.* Cancer Research, 2013.

186. Shyh-Chang, N., et al., *Influence of Threonine Metabolism on S-Adenosylmethionine and Histone Methylation.* Science, 2013. **339**(6116): p. 222-226.

187. Ulanovskaya, O.A., A.M. Zuhl, and B.F. Cravatt, *NNMT promotes epigenetic remodeling in cancer by creating a metabolic methylation sink.* Nat Chem Biol, 2013. **9**(5): p. 300-6.

188. Sun, Z., et al., *Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing.* PLoS One, 2011. **6**(2): p. e17490.

189. Wang, Y., et al., *Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells.* Proteomics, 2011. **11**(10): p. 2019-26.

190. Strande, V., et al., *The proteome of the human breast cancer cell line MDA-MB-231: Analysis by LTQ-Orbitrap mass spectrometry.* Proteomics Clin Appl, 2009. **3**(1): p. 41-50.

191. Jain, M., et al., *Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation.* Science, 2012. **336**(6084): p. 1040-4.

192. Marcotte, R., et al., *Essential gene profiles in breast, pancreatic, and ovarian cancer cells.* Cancer Discov, 2012. **2**(2): p. 172-89.

193. Jerby, L., et al., *Metabolic associations of reduced proliferation and oxidative stress in advanced breast cancer.* Cancer Res, 2012. **72**(22): p. 5712-20.

194. Isaacs, J.S., et al., *HIF overexpression correlates with biallelic loss of fumarate hydratase in renal cancer: novel role of fumarate in regulation of HIF stability.* Cancer Cell, 2005. **8**(2): p. 143-53.

195. Yogev, O., et al., *Fumarase: a mitochondrial metabolic enzyme and a cytosolic/nuclear component of the DNA damage response.* PLoS Biol, 2010. **8**(3): p. e1000328.

196. Ritterson Lew, C. and D.R. Tolan, *Targeting of several glycolytic enzymes using RNA interference reveals aldolase affects cancer cell proliferation through a non-glycolytic mechanism.* J Biol Chem, 2012. **287**(51): p. 42554-63.

197. Yang, W., et al., *Nuclear PKM2 regulates beta-catenin transactivation upon EGFR activation.* Nature, 2011. **480**(7375): p. 118-22.

198. Christofk, H.R., et al., *The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth.* Nature, 2008. **452**(7184): p. 230-3.

199. Schomburg, I., et al., *BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA.* Nucleic Acids Res, 2013. **41**(Database issue): p. D764-72.
200. Desvergne, B., L. Michalik, and W. Wahli, *Transcriptional regulation of metabolism.* Physiol Rev, 2006. **86**(2): p. 465-514.
201. Patil, K.R. and J. Nielsen, *Uncovering transcriptional regulation of metabolism by using metabolic network topology.* Proc Natl Acad Sci U S A, 2005. **102**(8): p. 2685-9.
202. Zelezniak, A., et al., *Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes.* PLoS Comput Biol, 2010. **6**(4): p. e1000729.
203. Masaki, H., Y. Okano, and H. Sakurai, *Differential role of catalase and glutathione peroxidase in cultured human fibroblasts under exposure of H2O2 or ultraviolet B light.* Arch Dermatol Res, 1998. **290**(3): p. 113-8.
204. Reuter, S., et al., *Oxidative stress, inflammation, and cancer: how are they linked?* Free Radic Biol Med, 2010. **49**(11): p. 1603-16.
205. Glorieux, C., et al., *Catalase overexpression in mammary cancer cells leads to a less aggressive phenotype and an altered response to chemotherapy.* Biochem Pharmacol, 2011. **82**(10): p. 1384-90.
206. Favaro, E., et al., *Glucose utilization via glycogen phosphorylase sustains proliferation and prevents premature senescence in cancer cells.* Cell Metab, 2012. **16**(6): p. 751-64.
207. Maddocks, O.D.K., et al., *Serine starvation induces stress and p53-dependent metabolic remodelling in cancer cells.* Nature, 2013. **493**(7433): p. 542-546.
208. Hatzivassiliou, G., et al., *ATP citrate lyase inhibition can suppress tumor cell growth.* Cancer Cell, 2005. **8**(4): p. 311-321.
209. Chiarugi, A., et al., *The NAD metabolome--a key determinant of cancer cell biology.* Nat Rev Cancer, 2012. **12**(11): p. 741-52.
210. Sahm, F., et al., *The endogenous tryptophan metabolite and NAD+ precursor quinolinic acid confers resistance of gliomas to oxidative stress.* Cancer Res, 2013. **73**(11): p. 3225-34.
211. Feun, L., et al., *Arginine deprivation as a targeted therapy for cancer.* Curr Pharm Des, 2008. **14**(11): p. 1049-57.